INTRODUCTION TO THE USE OF STANDARD TESTS

A BRIEF MANUAL IN THE USE OF TESTS OF BOTH ABILITY AND ACHIEVEMENT IN THE SCHOOL SUBJECTS

* * *

By SIDNEY L. PRESSEY, Ph.D.

Assistant Professor of Psychology
The Ohio State University

AND

LUELLA COLE PRESSEY, Ph.D.

Instructor in Psychology
The Ohio State
University

GEORGE G. HARRAP & CO. LTD.
LONDON CALCUTTA SYDNEY

THE PLIMPTON PRESS · NORWOOD · MASS · U · S · A

CONTENTS

PART FOUR — IMPORTANT GENERAL PRINCIPLES REGARDING TESTS

INTRODUCTION TO THE USE OF STANDARD TESTS

INTRODUCTORY

NO longer ago than 1908 appeared the first systematically organized form of the Binet Scale. In 1910 was published the first of the modern scales for the measurement of classroom products — Thorndike's Scale for Measuring the Quality of Handwriting. In 1915 Otis devised the first group scale for measuring general intelligence, embodying the principles later used in the army group tests. These dates may be taken as marking the real beginnings of the present efforts to measure mental ability and school attainment, though of course there had been scattered work in both lines before these years.[1]

The measurement movement, therefore, is comparatively new; in fact, the most important developments have come during the past few years. Particularly did the extensive psychological work done in the army during 1917 and 1918 give a new impetus to the movement, and establish new methods in testing. The testing of over one million draftees, and the use of test results in placing these men where they would be of the greatest value to the service, demonstrated the usefulness of these methods in a striking and dramatic way. As a result of the army work, and of continued

[1] Thus Rice's report on spelling in 1897 may be considered the first effort, in this country, to investigate empirically the product of teaching, and Cattell's paper in 1890, dealing with his tests of Columbia freshmen, the beginning of measurements of ability.

trial of tests in the schools, it may be said that the usefulness of the test method is now an established fact. And tests are being used to an extraordinary extent, both in the public schools and in the selection of employees by business houses. A large number of excellent tests are now available; methods are gradually becoming standardized.

Now that tests are being used very extensively, by teachers and others not especially trained in their use, there is needed a simple and direct manual, covering tests of both ability and achievement, to make clear to such workers the fundamental facts with regard to tests, the handling of test results, and the significance to be attached to the results of testing. Particularly there is need of a systematic treatment of the ways in which test results may be used in dealing with school problems. The present brief text aims to fill, in some measure, the need for such a manual.

The effort has been first of all (1) to give a clear discussion of the nature of tests, the problems that may be dealt with profitably by means of tests, simple methods for the handling of test results, and common mistakes to be avoided. Every effort has been made to keep the treatment non-technical, and to make it of the most practical character. In the second part of the manual (2) certain of the best and most representative tests in the various school subjects are presented. This presentation has been made brief, since, with the present activity in this field, improved instruments are constantly appearing and present scales and tests will become out of date. Therefore attention has been concentrated rather upon the fundamental problems of

measurement in the various fields, and present tests have been used as illustrative material. In the third part (3) tests of ability, particularly tests of general intelligence, have been discussed in the same way, with special emphasis on the practical uses of test results. Finally (4) certain general principles with regard to the way in which test work should be organized, and related to practical problems, have been emphasized.

The manual is thus intended as an introductory handbook in the use of tests. The effort has been, throughout, to suit the book to the needs of the busy teacher, principal, or superintendent.

PART ONE

HOW TO USE TESTS

CHAPTER ONE

What Is a Test?

BEFORE the appearance of the standard tests and scales with which this book is to deal, "test" was synonymous with "examination" or "written review." A test consisted of a number of questions selected by the teacher, and covering what she thought to be the important points of the subject studied. But with the introduction of the standard scales the word "test" has taken on a new meaning — a meaning which, superficially, appears very thoroughly to differentiate it from the ordinary school examination. Many teachers and administrators are still skeptical of the value of the new "tests" advocated by educationalists and psychologists. These practical schoolmen are skeptical not only of the practical value of the results obtained by use of these tests, but of the essential soundness of the present test movement. There is a feeling that testing is a time-consuming, artificial, and one-sided way of obtaining information which the teacher may acquire much better by her own observation of her pupils and their work. These circumstances make it necessary that the exact nature of the tests and the advantages of the test method should be thoroughly understood. It will be found the fundamental contention of this book that tests, as the term is now used, are not merely a very special method to be used only on special occasions. Instead, the new tests are the result of an effort to avoid certain limitations and failings of the usual written examinations. And it will be argued that, far from being a special method to be used only rarely,

tests should be part of the equipment of every school, because tests will do much that has heretofore been done not at all or not so well by the usual examinations given by teachers or principal.

Three distinctive features of the "test," as compared with the usual school examination, will be pointed out.

1. CHOICE OF SUBJECT MATTER OF STANDARD TESTS

A good test covers only the really important points of a subject. A teacher intends to cover only important points, but no one person can have an unbiased opinion as to what is important. The maker of a test consults as many textbooks as he can find, talks with as many teachers of the subject as possible, corresponds with advanced students and experts in the subject under consideration, and so comes to a conclusion as to what points, in the total amount of subject matter, are of greatest importance. He then discards those questions that seem of relatively little significance and keeps only those which deal with basic facts. The subject matter of the tests is, then, far more carefully considered than are the questions of a school examination. The writers themselves have been guilty of making up the questions for an ordinary classroom examination while riding in the street car on the way to class; and they know some teachers who announce there will be a written lesson and then turn to the blackboard and write down such questions as occur to them while writing. It is evident that even the careful reflection of one person — and such reflection is not infrequently missing — cannot produce questions of so consistently

important a nature as can the thoughtful consideration of the problem by many people.

For instance, the Hahn-Lackey Geography Scale consists of a series of questions very carefully chosen. First, questions common to six modern textbooks were selected; only the essentials of the subject matter were thus included. This list of questions was also checked by study of the curricula worked out by special students of geography teaching in the public schools. Then the wording of these questions was very carefully examined to eliminate expressions that were over-technical. And the exercises were revised so that they contained about an equal number of "memory" and "thought" questions. It should be obvious that questions thus carefully selected are much more valuable than questions which might be chosen by a single teacher, no matter how careful she might be in making up her examination.

The best tests are based on very careful research as to the fundamental objectives in the subject concerned, and the material is selected with reference to its importance for these objectives. Thus, before constructing the Ayres Spelling Scale its author determined, by very careful and extended investigation, the words actually used most commonly by children and adults, in written work. The scale was then made to include only the one thousand words used most frequently. Clearly these are the one thousand words which it is most necessary that children should learn to spell.

Moreover, not only is the importance of the questions in a standard test determined, but also their relative difficulty. This is evidently of decided importance. The questions of a test should be neither too easy for the most capable children in the class nor too hard for the poorest children. Preferably the questions should be arranged in order from the easiest to the hardest.

And the increases in difficulty should be regular. That is, Question 6 in the test should be as much harder than Question 5 as Question 5 is harder than Question 4. Actual experimentation has shown that it is practically impossible for a teacher to estimate the difficulty of questions; a question which she thinks easy, and to which she assigns little credit, may really be hard for the pupils.[1] The difficulty of each question in a standard test has been carefully determined by trial with a large number of school children. The test thus contains materials suitable for all degrees of ability in the grades to which it is applicable; there are some questions easy enough for the poorest pupils and some questions hard enough to try the mettle of even the best students. And the regular increases in difficulty make the test into a "scale" in which the units are, in a very important way, equal. And we can say that a child who scores 5 on the scale is as much better than the child who scores 4 as he is poorer than the child who scores 6; the child with the score of 5 does work as much more difficult than 4 as it is easier than 6.

> Thus, Monroe's Silent Reading Tests begin with items which are very easy and progress by comparatively even steps to items of decided hardness, for the children to whom the test is supposed to be given. And a child is credited more for doing a hard item than for doing an easy one.

A "test" is thus to be distinguished from the usual school examination first of all by the extreme care used in the selection of materials for the test. The questions

[1] Comin, Robert, "Teachers' Estimates of the Ability of Pupils." *School and Society*, Vol. 3, page 67, 1916.

for a test are selected only after very careful study of textbooks and consultation with authorities, in order to get a consensus of teaching practice and of expert opinion regarding the subject matter to be covered. Where possible, selection of material is based upon searching analysis and extended investigation regarding ultimate objectives in the teaching of the subject in question. And the difficulty of the problems to be used is very carefully considered, so that either a regular progression in difficulty, or a series of problems of equal difficulty, may be secured.

2. PROCEDURE IN GIVING, TAKING, AND SCORING THE "TEST"

The way in which the test should be given is specified in detail. Particularly in the way they give examinations, teachers differ strikingly. If the reader wishes to experiment with this point, let him write out a series of questions in some subject, and then ask several teachers of the same grade how much time they would give their pupils for answering these questions, and what their procedure would be. He will find that some teachers will allow twice as much time as other teachers will consider sufficient. He will also find that some teachers will allow the children to ask questions if they do not understand what they are to write, whereas other teachers will allow no comments. If he were to watch the various teachers while they gave the examination, he would find that their methods were very different, some reading the directions only once and perhaps hurriedly, while others went over them in detail two or three times. It is evident that the results

from the same set of questions from two different rooms cannot be compared if the children in one room are given more time than those in another, or if the teacher in one room gives directions for answering the questions that differ in any essential way from those given by the teacher of the other room. Every one who has taught knows that classes will do vastly different work if told to "work rapidly" from what they will do if told to be "as accurate as possible." And any encouragement given by the teacher, even if only a request that they do their best, will influence the work done by the children. Consequently, it is obviously impossible to compare the papers from children in different rooms, schools, or cities unless the method of giving the tests — the procedure — is identical for all classes tested. Many a teacher wishes to know whether or not her class ranks as well in a subject — say arithmetic — as the class taught by another teacher in another building. But a comparison of the two classes is not justified unless both the questions asked and the procedure used in the giving are identical. The standard tests avoid these difficulties. The directions for giving a standard test are to be read verbatim, as prescribed by the author of the tests. With the best tests, the directions are printed on the test blank, so that failure on the part of the teacher to read the directions distinctly will not influence the results. The conditions for testing are thus rendered perfectly constant — or identical — from class to class, no matter where or by whom the test is given.

Thus the directions for the Courtis Arithmetic Test, Series B, are printed just above the examples, and consist of the following explanations: "You will be given eight

minutes in which to find the answers to as many of these addition examples as possible. Write the answers on this paper directly under the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples." In giving the test the teacher has the pupils read these directions aloud with her. And, since the directions are on the same page with the test itself, the children can always glance back at the instructions if they find themselves in doubt as to what to do. So each child has the same opportunity to understand what is to be done and works with exactly the same directions — directions which are presented in exactly the same manner to every other child who is given these tests, whether in the same room or school or not. And the time allowed each class is always the same.

There is also careful control of all factors which might, after the directions were given, affect a child's work on the tests. Thus a history test should require little if any writing on the part of the pupil. The reason for this is simple. If there is much writing, the child who writes rapidly will get more questions done and will, therefore, have a greater chance to get a larger number of questions right than another child who knows just as much about history, but writes so slowly that he has time to answer but a few questions. In any case, the ability to express oneself quickly and adequately in written English is an ability which varies enormously from child to child; and, if a teacher is trying to measure a child's knowledge of historical facts, she does not wish to have the results obscured by such accidental factors as speed of writing. Every teacher has some children who never do themselves credit on an examina-

tion because they cannot write down their ideas readily. In order to avoid such difficulties the best "tests" require little or no writing from the child. Instead, the child underlines words, crosses out words, or does something of that sort which will not give any considerable advantage to the child who writes rapidly or the child who expresses himself as easily on paper as in a recitation.

Thus a vocabulary test included in the Haggerty Examination "Sigma 3" includes such lists as:

calm (quiet, sleepy, night, restful)

cupola (church, high, schoolhouse, rounded dome)

and the children are told to "draw a line under the best definition for each word." If the children were simply given the words "calm" and "cupola," and were asked to write a definition, the child who wrote the most rapidly would get the most answers done within the time limit, though he might not know as many words as a slower writer. Further, if the child who knew the meanings of the words most exactly defined them most fully, he would define fewer words within the time limit than the child who simply gave a rough definition for each word and then rushed on to the next. By offering a choice of answers, among which the child is to indicate the correct one by underlining it, all these various difficulties are largely done away with.

The scoring of a good "test" is quite as unmistakable and clearly defined as the procedure in giving. In the newest tests answers are not written; certain marks — underlining, crosses, or checks — serve to indicate the correct reply. At most, only a word or two, or a number, serves for an answer. There is thus little opportunity for a mistake in grading, or for doubtful responses on the part of the child. Either the correct word or

number or mark appears where it should, or it does not; the answers are always unmistakably right or wrong. The great exactness and objectivity thus possible in grading tests should be compared with the very loose methods usually employed in marking ordinary examination papers. Suppose a teacher of history, in a junior high school, gives an ordinary examination to her class. The mark a given boy receives is almost certain to be dependent not entirely upon the amount of history he knows; the neatness of his paper, the legibility of his handwriting, his facility in written expression, are almost certain to influence the teacher's mark to some extent. Moreover (a very important point indeed), the mark assigned by the teacher is also dependent upon her standards as to what should be required of a class, her judgment as to the comparative importance of various points, and her conception as to the worth of various types of answer. In spite of the best of intentions the mark may be influenced somewhat by her mood at the time, and perhaps by her attitude toward the boy in question. In contrast the score which a boy obtains on a modern "test" is always fair and impartial, and uninfluenced by the special factors just mentioned; the score is thoroughly objective.

Any teacher who doubts what has just been said about the unreliability of marks should try two or three simple experiments. She should, for instance, give an examination and mark the papers, keeping the marks on a separate sheet, — and then put these papers away in a safe place. After the lapse of a month or so she should grade these papers again. Many of the second marks will be found surprisingly different from the marks first assigned. And, if a second teacher is asked to grade these same

papers, still more surprising discrepancies will appear. The matter has, in fact, been very carefully investigated.[1] And it has been proved that a group of mathematics teachers may vary all the way from 28 to 92 in the marks they assign to the same final examination paper in geometry. It should be obvious that no such variations are possible in scoring a good test. Thus, in grading the vocabulary test mentioned above there is almost never any question as to whether or not a given reply is correct; either the correct word is underlined or it is not.

To summarize: the second great advantage which a "test" has over a school examination is that the procedure in giving, the nature of the child's answers in taking, and the process of scoring are very carefully worked out and can be kept constant from one class to another. As a result, it is possible to compare the work of one class with the work of another, with a definiteness which was undreamed of a few years ago. Norms, or "average scores," based on the work of several thousand children of a given age or grade are available, for the best tests. So a teacher may compare the standing of her pupils not only with the records made by other classes in her own school or city but with the general average or "standard" obtained from schools all over the country. The teacher of a little country school can find out whether or not her four sixth-grade pupils can do, in arithmetic, or spelling, or history, what has been decided upon as "average" for pupils of that grade in school systems generally. From the point of view of comparability of results the standard test is unique.

[1] Starch, Daniel, *Educational Psychology*, pages 426–438.

3. SAVING OF TIME AND LABOR

The statement that the test method saves time and labor will be immediately challenged by many teachers who may have slaved for hours over the scoring and interpretation of some awkwardly constructed test. Certainly, not all the tests now on the market are easily given and scored. However, it may be said shortly that *a good test should save time and labor.*

Consider, for instance, the questions in American history which appear below:

The people who settled Plymouth were: Dutch English French German.
The Quakers settled in: Quebec Maryland Vermont Pennsylvania.
The Spaniards explored America chiefly to obtain: furs gold trade land.

This eighth-grade test contains 40 such questions. The children are given 6 minutes in which to answer these 40 questions; usually more than half of an eighth-grade class will complete the test. How many questions of the usual school examination could a class of children answer in 6 minutes? Surely, even with brief "fact" questions, less than half that number. But the great saving of time is in the scoring. It is possible for a teacher to score the 40 questions of the history test just referred to at the rate of almost two papers per minute. To realize the saving of time one should think of the weary task which a teacher would have before her, in grading a 40-question examination from a good-sized class. To score the test requires not over 30 minutes for a class of 35 pupils. Examination papers from the same class could hardly be graded in four times that period. And there still remains to be taken into account the fact

that the test comes to the teacher with the questions all prepared, and everything ready for use, whereas the examination questions for a written lesson must be carefully made out by her. Surely it would seem not unreasonable to claim that tests may soon lighten, very materially, the drudgery of teaching.

It may be argued, in opposition to the above point, that some of the group scales for measuring general intelligence are hardly easy to score. But it must be remembered that these more elaborate scales really include an extraordinary amount of matter. Thus the army scale Alpha is not altogether easy to score. But it contains a total of 212 questions; and it yields information of very great general significance. When these facts are taken into account, one realizes that the "time-cost" is relatively slight.

To summarize, then, these three facts are of greatest importance to understand, in appreciating the difference between a "test" and the usual school examination: (1) The subject matter of a test is selected with extreme care as matter stressed in the best teaching practice, and of the most practical worth; and the difficulty of this material is determined. (2) Methods of giving, taking, and scoring are very carefully worked out and standardized, so that results are of a clean-cut significance, and wide comparisons from class to class, school to school, or city to city, are possible. (3) A good test is a great saver of time and work, and permits one to obtain a really extraordinary amount of information about a class, school, or school system with an astoundingly small expenditure of time and energy. Surely an educational instrument to be distinguished by such characteristics is neither impractical nor of narrowly

limited and only occasional usefulness. And the question at once arises as to what, specifically, tests may do, in contributing to the solution of educational problems.

In this matter one possible misunderstanding, perhaps, should be cleared away at once. Throughout the present chapter the writers have contrasted tests with the usual school examinations. It may possibly be inferred that they believe examinations set by the teacher should, as soon as possible, be entirely abolished — tests being always used instead. Such an inference would be decidedly unfortunate. The examination set by the teacher has its own distinctive service to perform in keeping her in more intimate contact with the work of her class than is possible otherwise, in enabling her to follow special features in her teaching, in giving opportunity for more spontaneous and original work than is called for in the usual test. Tests should not completely take the place of the usual examination;[1] they should, rather, be used to render service which examinations can never render. It is especially with these distinctive contributions which the tests can make, in dealing with school problems, that the next chapter is concerned.

[1] It should perhaps be mentioned, in this connection, that tests are often of great service to the teacher when she is formulating an examination or quiz; in fact, such scales as the Hahn-Lackey Geography Scale are simply mines of material to which the teacher may go for such purposes. And the best tests serve not merely as measuring instruments. Certain indirect values are quite as important. The test formulates in very concrete fashion the minimal essentials in a subject, makes clear the teaching objectives, and introduces the teacher to scientific methods which she will find valuable in all her work.

CHAPTER TWO

School Problems and Tests

At any convention, where school superintendents, supervisors, or teachers are gathered, tests are one of the chief topics of discussion. The use of test materials during the past few years has been amazing. Millions of copies of some of the best-known tests have been used. It is quite evident that the school people of the country are becoming more and more interested in "tests." One naturally asks what is being done with the results of all this work; surely the tests must contribute to the solving of teaching and administrative problems, if they are to continue long to be thus extensively employed. However, it often appears as though tests were given for no other reason than because every one else was doing so. The writers have known several school superintendents who bought — at considerable expense — blanks for a survey of their system, made the survey at the cost of much time and effort on the part of all concerned, and then — filed the papers and made no use whatever of the results. This last procedure is far too common. In fact, it is no exaggeration to say that only a small percentage of the total number of tests given in a year are made to yield anything like the help, to the teachers and school officials of the schools concerned, that they could give.

Always it should be kept in mind that testing for the mere sake of testing is an exceedingly unwise as well as expensive operation. The person who sets out to give a test should always have *some* reason for doing so — should have some practical problem in mind. He

should always make *some* use of his data. The present chapter will aim to point out some of the problems of teaching, supervision, and administration to the solution of which tests may be expected to contribute.

For purposes of clearness and convenience the distinctive problems of (1) the teacher, (2) the supervisor, and (3) the superintendent will be discussed separately. The major educational problems, of course, affect all three groups; in so far the distinction is somewhat artificial, and the discussion involves some duplication. However, the situation faced by each one of these persons is different. The problems of the teacher have to do largely with instruction; and the teacher is interested in the use of tests to improve instruction. A teacher is concerned, in the last analysis, with the handling of the individual pupil in her class. The problems of the supervisor or principal have to do largely with the coördination of the work of various classes and teachers, and the judging of the work of the teaching force. The supervisor deals, not with the individual child, but particularly with classes or schools; she is interested in the measurement of groups. The problems of the administration have to do with larger matters of school policy. The superintendent is interested in the comparative accomplishment in various schools of the same general type, in general questions regarding the adequacy of the curriculum and the efficiency of the system, in the value of various methods of organization, and so on. To all of these problems the use of tests may reasonably be expected to contribute; but the special values, and limitations, of tests must be constantly kept in mind. The special

problems of the teacher, the supervisor or principal, and the superintendent will be taken up in order.

1. TESTS AND THE PROBLEMS OF THE TEACHER

(*a*) *Problems involving the measurement of general mental ability.* As has been said, it is the major duty of the teacher to instruct. Consequently, she will wish any tests she uses to assist her in improving her instruction. First of all, she is interested in anything which may throw light upon the abilities of the children she is to teach. She should know as much as possible about the abilities of each child in her room; for, other things being equal, the teacher who knows her children thoroughly will be a better teacher than one who is content to think that all her pupils are about alike, and so treats them as a group instead of as individuals. So a teacher may well begin her use of tests by employing them to give her a knowledge of the capacity of each child in her class. During the first week or two of school she should give her class a group scale for measuring general ability.

She will probably find that even in a small class the children differ strikingly as regards ability. If possible it would probably be worth while to divide her class into two or more sections on the basis of ability; the brighter children should be put into a "fast" section and the duller into a "slow" section. Very likely the teacher is accustomed to dividing her class into such sections. If she uses the group scale to help her in this classification, she will find her divisions distinctly more accurate than would be the case otherwise. Particularly, the tests will give her information permitting her to make such

section divisions shortly after the beginning of school without waiting for extended acquaintance with the children, and these divisions will be far more permanent than any first divisions of her own making. Once the teacher has her class satisfactorily sectioned, she will find that she can adapt her instruction to the ability of each group, and so approximate the needs of each child much more readily than she could do otherwise.

However, adjustment to differences in ability should not end with section division. A special effort should be made to discover any exceptional children and to make whatever adjustments for them may be possible or necessary. If a teacher discovers in her class a remarkably able child, she owes it to him to give him such special help as will enable him to pass through the elementary and grammar school work at a rate commensurate with his ability; thus time will be saved the child, and he will be given sufficient extra work to keep him occupied and interested. If the class includes one or two very stupid children, the teacher owes it to them to emphasize such elements in the regular curriculum as they can make the most of. And she owes it to herself not to become discouraged if these subnormal pupils do not learn readily, in spite of her best efforts.

In fact, it is coming to be accepted as a general principle that a teacher should obtain from each child achievement in proportion to the child's ability. The brilliant child should do brilliant work and the feeble-minded child should do "feeble-minded" work — and each child should, from one point of view, receive as much credit for his effort as another. However, it is obvious that a teacher cannot hold each child to work

corresponding to his ability unless she first knows what that ability is. Results obtained with a good group test of intelligence at the beginning of the year make an admirable basis for such efforts to individualize instruction.

(*b*) *Previous preparation in the several school subjects.* The teacher will wish, at the beginning of the school year, to obtain information not only with regard to the general ability of her new pupils, however. She will wish also definite facts with regard to their previous preparation in the subjects in which she is to instruct them. For this purpose she should use general tests in the school subjects — in arithmetic, reading, and so on. She will find that she can obtain an idea of the previous preparation of her children much more easily and quickly by means of such tests than by examinations or class work. The Courtis Arithmetic Tests, for instance, will show her at once the ability of her class in the fundamental operations of arithmetic. A little time, during the first two or three weeks of the semester, devoted to testing will indicate clearly the general standing of the class in their work to date, and show the teacher where she should begin her own instruction. The tests will also discover any children who are conspicuously lacking in previous preparation, or otherwise so strikingly different from the rest of the class in previous training that it is advisable to give them special attention or (possibly) to place them in a different grade or class.

(*c*) *Diagnosis of special weaknesses.* So the teacher will wish to learn first of all, by means of the general tests, the standing of the various members of the class

in each subject. However, as the work of the year develops she will need more detailed information than this. If the class is poor in the fundamental operations she will wish to know on which particular points it is weakest, so that she may emphasize these points and remedy the difficulties. Perhaps the Courtis tests show the class to be particularly weak in addition. This evidently tells the teacher something with regard to the weakness of her pupils; but she needs more specific information than this. To obtain this information she should give "diagnostic" tests in arithmetic. These may reveal that the class is poor in addition primarily because of failure to "carry" properly. The teacher then at once knows just what should be emphasized in her teaching — where explanation and drill must be put. Or it may be that two or three children only show this failure in addition; it then appears that for these children in particular "carrying" must be stressed. In either instance, a diagnostic test gives the teacher information which a general test could not furnish. And it gives information which she could not have obtained as easily or clearly otherwise. In time, of course, the teacher would have found out that the class was weak in "carrying" — even though no test had been used. But the fact would not have been brought out in such clean-cut fashion; nor would the information have been obtained at once, at the beginning of the school year, while there was still a maximum amount of time in which to remedy the difficulty. If only two or three of the children are weak in some such special capacity, the teacher may never discover just what is the matter with their work. By means of the diagnostic test she

can at once put her finger on their difficulty, and at once emphasize exactly those points most needing attention.

(*d*) *Remedial instruction.* However, tests will not only indicate what specific points need to be stressed and diagnose the instructional problem. In certain subjects there are now available test materials which will be of very great assistance to the teacher in her remedial instruction and will almost automatically inform her, from time to time, of the gains of her pupils. Thus if a child is weak in "carrying" he may be given practice materials involving "carrying"; each day the child works on these materials, checks his answers, and records his progress so that she may observe it. And, at the end of a certain period, the teacher gives a new test which informs her still more definitely of the improvement made and the weaknesses of each child at this period. These systematic practice materials are, in some respects, the finest results of work in the field of educational measurement.

The teacher, then, may use tests in the solution of four types of problems. She may use (*a*) tests of general mental ability in order to find out the general ability of each child in her class, so that she may adapt instruction to the capacity of each child. She may use (*b*) general measures in the school subjects in order to determine the previous preparation of the class and their general achievement to date. She may give (*c*) diagnostic tests in the school subjects in order to find out the specific weaknesses which it should be her special effort to remedy. Finally (*d*) systematic practice exercises will assist and guide her in any needed remedial instruction.

2. TESTS AND THE PROBLEMS OF THE SUPERVISOR

(*a*) *Progress of pupils in school subjects.* A supervisor is, by definition, a person who oversees the work of others. As such she is responsible for the methods used by those under her and for the coördination of effort on the part of those whom she supervises. The supervisor, then, needs always to know what the various classes are doing in the school subjects. In the course of time a teacher can, because of her constant association with her children, come to form a very fair estimate of their needs and of their progress without the use of any special devices; but the supervisor has no such opportunity for detailed investigation. She must have some way of following the progress of a very large number of children. Evidently, tests are of distinctive value to the supervisor for this purpose.

(*b*) *Comparison between classes and schools.* The supervisor should know also the relative progress made by various classes and schools in the different school subjects. By the use of tests this comparison can be made in very definite numerical terms, and (no unimportant point) can readily be obtained in such a way that there can be no question about the reliability or impartiality of the findings. Such a measure takes supervision, at one stroke, from the realm of opinion and guesswork to that of definite fact. And when such results *are* obtained, it is very easily demonstrated that schools and teachers vary strikingly in the results they are obtaining. It is the supervisor's duty to discover such inequalities and then, as far as possible, to remedy them.

(*c*) *Determination of the causes of differences between classes.* A number of causes may operate to produce differences between classes; and, clearly, the supervisor must find the cause before she can advise as to the remedy. First to be considered is the possibility that the work in one school may be poor because the children in the school are unusually dull and consequently, even with the best of instruction, do not learn readily. Evidently if a supervisor is to deal adequately with the situation she must in some way discover the nature of the "pupil material" with which each school has to deal. This can be done by means of group tests of intelligence. One of the striking results of recent research has been the demonstration of large differences, from school to school, in the ability of the children attending these schools. And a supervisor may do a large injustice to the teachers in a "poor" district of a city if she fails to take account of this factor and attributes to poor teaching what is really due to the poor innate capacity of the children who are being taught.

The supervisor should also, in this connection, study the "age-grade" situation; it may be that a school showing a high achievement has obtained this apparent excellence by retarding the children and keeping all the dull children in the lowest grades. These points will be returned to in a later chapter.

(*d*) *The adequacy of teaching methods.* Finally (and not until these points just mentioned have been covered) the supervisor should investigate the teaching methods employed in the various classes. In this investigation she will find tests of very great value, both in a first appraisal of the teacher's work and in

measuring the effect of any suggestions which she may offer. And she may carry as far as seems advisable, by means of diagnostic tests, her analysis of each teacher's difficulties, and may make systematic use of practice materials or otherwise carry out controlled efforts at systematic improvement.

A supervisor — or a principal — should be more than merely a critical adviser. She should forge ahead in the matter of teaching method. She should make experiments in this or that school or determine the adequacy of various methods of instruction. No single teacher can do this satisfactorily; in fact, a teacher is hardly in a position to deviate very extensively from the specified material and methods. Such experimentation must, then, be carried on largely by the supervisor. And in such experiments tests are an invaluable aid. If the conditions are properly controlled, it is usually possible to come to a very definite conclusion concerning the excellence of this or that method of teaching.

To summarize, then: for the supervisor, tests are particularly of use in four fairly distinct ways: (*a*) She may use tests in the subject she supervises to follow the progress of the different classes and (*b*) to discover any unevennesses in the results of the teaching. (*c*) She may, further, use them to find out the causes for any inequalities, and to make sure that they are remedied, if this is at all possible. (*d*) She may use tests as an aid in investigation of the adequacy of various teaching methods.

3. TESTS AND THE PROBLEMS OF THE SUPERINTENDENT

The superintendent is the responsible executive in charge of the operation of a school system. As such, any means which will keep him in close touch with the condition of that system and enable him accurately to direct and appraise the work of those under him will evidently be of value. To a certain extent his knowledge of what is going on in his system will come through his supervisors, or through the principals acting in a supervisory capacity. However, tests may be used by the superintendent in certain rather distinctive ways, in dealing with problems chiefly administrative in character.

(*a*) *The need for a measure of efficiency.* It is very necessary that the administration shall know the efficiency of the teaching staff in an impartial and exact fashion. For this purpose tests are of great value. It should be obvious, however, that tests ought not to serve primarily, in a school system, as a means of obtaining evidence on the basis of which a teacher may be "fired." In certain school systems a very unfortunate attitude of hostility to tests has been developed among the teachers because of their use chiefly in this critical fashion. The writers would almost say that if the tests are to be used only as evidence against the teachers they had best not be used at all; tests should serve primarily as an *aid* to the teachers — and be used by the supervisors to obtain a basis for helpful suggestions rather than for destructive criticisms. Nevertheless, if tests are used as widely and extensively as they should be, they will not only aid teacher and

supervisor, but also serve as one more means by which the superintendent may be kept informed as to how things are progressing in his system. This information, however, should not deal solely, or even primarily, with the teaching efficiency of the individual teachers. The teacher is not by any means the only factor in the educational situation — though she seems to be the one who is usually blamed for any shortcomings! And to the study of many of these other factors tests may contribute.

(*b*) *Determination of accuracy of grade placement.* For instance, how about promotion policy? Perhaps the superintendent is making a special effort to cut down the retardation, and is asking his teachers to fail fewer pupils than in previous years. However, he does not wish the teachers simply to pass every child more or less indiscriminately. Presumably he wishes those children promoted to the next grade who are best able to do the work of that grade. Particularly (if he wishes to follow out the newer ideas with regard to facilitation of progress through the grades for the unusually bright children) he will wish to make certain that the more capable children in his system are being moved through school rapidly, and not being held back to waste their time in grades where the work is too easy for them. In investigating this matter group tests of intelligence will give very valuable information. If the superintendent surveys his system, using some one of the group scales for measuring general intelligence now available, and then tabulates the results by grade, he has before him on the tabulation sheet, very prettily exhibited, the accuracy of grade placement in the various schools.

In schools where the promotion has been indiscriminate he will find children of all degrees of ability in each and every grade. Especially (a most unfortunate situation) he will find no effort at double promotion, or other special opportunity for those children who are unusually able. And he can very easily obtain definite measures of the extent to which a given school has failed in wise promotion; he can, for instance, count up the number of children from each school who show ability two grades in advance of the grade in which they are now placed. Superintendents and teachers are generally familiar with age-grade tables. Much more worthy of study is a "mental" age-grade table. Such a table may yield a superintendent or principal very illuminating information with regard to the accuracy with which children have been assigned to the various grades or sections.

(*c*) *Tests and educational reorganization.* If such a survey is made, by means of group tests of intelligence, other facts are likely to emerge, however. It will probably be discovered that the children in certain schools are duller than those in other schools. This fact will usually appear partly in promotion rates; there will be more failures in the school with the duller pupils. It will also appear in lower scores on tests in the school subjects. But tests of intelligence are needed to explain both these facts; otherwise there is likely to be much misunderstanding, and possible injustice to the teachers having the poorer "pupil material" to work with.

If the intelligence tests are given it will then be demonstrated, with a clearness impossible otherwise, that schools differ in the problems they have to meet.

And it will then be evident that, so far as possible, special adjustments should be made for these special problems. It is particularly in the neighborhood showing a great many dull children that ungraded classes, and vocational work for those who will probably leave school as soon as the law allows, are desirable. It is in those neighborhoods showing children of superior intelligence that college preparatory work is needed.

(*d*) *The need for convincing facts.* These facts are, perhaps, easily obtained simply by observation of the various neighborhoods. But the test results exhibit the situation much more incisively than would be possible otherwise. It is possible, for instance, to demonstrate the need of an ungraded class for backward children very clearly, if one has figures showing that 10 per cent of the children in a certain school rate as feebleminded or mentally backward. The point hardly needs further illustration. It should be obvious that tests can be used so as to bring out, with an exactness which is often remarkable, the specific problems of the different schools in a system. The superintendent may understand the situation well enough. But the tests are needed to demonstrate the matter to his board — perhaps to put the matter before the public. Thus a county superintendent may use tests in the fundamental subjects to demonstrate the inefficiency of the teaching in one-room country schools and the need for consolidation.[1]

(*e*) *Comparison of entire systems.* One more type of information may be obtained by means of tests for

[1] Yawberg, A. G., "Third Annual Bulletin and School Directory of Cuyahoga County (Ohio) School District."

the service of the superintendent — and information which, especially, the superintendent could not obtain in any other way. Tests give him a means for comparing the work of his system with that of other systems. By means of standard tests he can find out, very definitely, whether his children are as proficient as children of the same grade in other cities. He may compare the accuracy of grade placement in his system with promotion policies in other systems. He may study in detail the results of any changes which he may initiate in his system, and compare the results with the results of similar experiments elsewhere. Such comparisons — it is worth while repeating — are almost impossible without the use of tests. By means of tests they are very readily made.

To summarize, then, the ways in which the superintendent, or other administrative officers, may find tests of service: (*a*) Tests in the school subjects will keep the superintendent in close touch with the efficiency of teaching in his schools, and will assist him in forming an opinion about the work of his teachers. (*b*) Tests of intelligence will be of great value in formulating his judgments with regard to promotion policies and the accuracy of grade placement in the various schools, and will make clear to him the nature of the problems which the different schools have to face, in serving the children who come to them. (*c*) Tests of both ability and achievement will be of great service in analyzing the educational situation in a system and in making clear what remedial measures should be taken. (*d*) Test results constitute incontrovertible facts, so often needed by the superin-

tendent in a campaign of education of public opinion. Finally (*e*) the tests will make possible comparisons which would be impossible otherwise with other systems, and so lead to a clearer appraisal of the efficiency of his system as compared with work done elsewhere.

So much then for a brief discussion of the problems of the teacher, supervisor, and administrator in the solution of which tests are a distinct aid. It should be added, however, that these values can be indicated in only a very general way. Each system, each school, each class, has its own problems, in the solution of which tests will often aid. The teacher, principal, or superintendent concerned must, naturally, work out for himself the application of tests to these problems. And each test usually has certain special merits of its own which make it of special service in some special way. It will be the aim of the writers to touch upon these more specific problems, and to make clear the peculiar values of the different tests, in later chapters.

CHAPTER THREE

The Common Sense of Statistics[1]

PERHAPS there is no matter concerned with the use of tests which teachers find so difficult, in connection with which they make so many unnecessary mistakes, or about which there is so much misunderstanding, as "statistics." The fundamental difficulty seems to be in a misconception regarding the nature of statistical methods. The average teacher looks upon them as being based on abstruse mathematical theories which she cannot hope to understand. So instead of trying to appreciate the rationale of these procedures, she is continually searching for some rule or theorem to cover any difficulty she may meet. And instead of being a help, the methods become a burden. Nine tenths of all these difficulties would be eliminated if teachers would only realize that these methods are simply convenient ways of handling data and of making clear the facts which they are trying to investigate. The methods should never be used blindly, according to rule, but intelligently and with a full understanding of their very practical nature. Fundamentally they are simply refinements on ways of handling results which any teacher would naturally work out for herself, if called upon to go over a large number of test scores and find out the important facts regarding them. They are simply refined "common sense." And it is the aim of the present chapter to make clear the "common sense" of these methods, so that the teacher may use

[1] This chapter may well be read after Part III by those as yet totally unfamiliar with tests.

them in such a way as to be of the greatest possible service to her.

1. SIMPLE CALCULATIONS MADE BY ARRANGING PAPERS IN ORDER

Suppose a teacher gives a certain reading vocabulary test, containing 34 questions, to a 3A class of 35 pupils. She has no need of any complicated statistics in handling and interpreting her results. After scoring her papers she should simply arrange them in order from highest to lowest. The papers as thus arranged will tell her a great deal about her class. She will at once see who made the best and who the poorest score, and how far apart on the scale these extreme cases are. If she plans to divide her reading class into two sections, she will know forthwith what the test results suggest regarding the make-up of the sections. She may study the agreement between the test and her own judgment of the children by listing the children (before scoring the test papers) in order as she estimates their reading ability, and then comparing this list with the list made up from the test scores.

Finding the median.[1] The teacher may wish more than merely a list of scores, however; she may wish to summarize her results in some way. Suppose the scores made by her children ran, in order, as follows: 32, 25, 24, 24, 22, 21, 21, 21, 20, 20, 20, 19, 19, 19, 19, 18, 18, 17, 17, 17, 17, 17, 16, 16, 16, 16, 15, 15, 14, 14, 14, 9, 9, 8, 4. Perhaps the teacher's first thought will be to add together all these scores, divide by the number

[1] Attention is called to the Glossary at the end of the book.

of cases, and so find the arithmetic mean or "average." A much easier method, and one more applicable to test results, is to find the "median," or middle score. Evidently in the above group of 35 cases the middle score will be the 18th score. So the teacher should begin at either end of her pile of papers — or list of scores — and count in until she comes to the 18th paper or score. This middle score is, in the case given above, 17.

Suppose there had been 36 in the class instead of 35. Evidently there would be no middle case; instead the median should be considered halfway between the 18th and 19th cases. If the child who has been added to the class is dull, and scores 9, then both the 18th and 19th scores will be 17, and the median will still be 17. If this added child is bright and makes a score of 22, then the 18th score (counting from the bottom) will be 17 and the 19th score 18. The median may then be considered 17.5. The logic of the method should be evident. The teacher should not consider the above statements a "rule" for finding the median; rather, the method is merely a convenient way of determining the "central tendency" of the class. If the teacher wishes she may consider the lowest of the two middle scores, in the last special instance just mentioned, the median. This will make the median again 17, and will avoid fractions; in dealing only with a single class this is, perhaps, the best way. Either method is satisfactory. And authorities differ in the exact methods they employ; other methods will be discussed in the Appendix. If the directions accompanying a test specify the method to be used, the teacher should of course follow that method. The important point is not to become bewildered by such differences of opinion, but to understand the general logic of the method.

The teacher should not confine her attention to the median, however. She should note the extent to which

the children in her class differ from each other. The most natural thing to do is to note the distance from the lowest to the highest score. The lowest score, in the example given above, is 4 and the highest is 32; this means that the total "range" is 28 points. Of course, the position of the extreme cases is likely to be somewhat variable and unreliable and this statement of total range not altogether satisfactory; so too much should not be made of it. But it is always worth while to glance at the "range" of the cases to determine how different the children of the same class are from each other. The reason for this is fairly simple: the child who scored 32 points certainly has a reading vocabulary that would enable him to read material altogether beyond the capacity of the child whose score was 4. A consideration of range should result in any adjustment in teaching that appears necessary in adapting lessons to the ability of the children.

Comparing the results with the norms. So far the teacher has considered only the record of her own class. But she will certainly wish to compare her class with other classes, and especially with the "norms." Again the method is simple. The norms for a test are usually sent out with the test papers, so that the teacher may have them for reference. In the particular test used as an example, the norms are as follows (these "norms" are simply the median scores for each grade, obtained from testing a large number of children):

Grade	2B	2A	3B	3A	4B	4A
Median Score, or "Norms"	6	9	10	14	21	26

The norm sheet shows a median score, for the 3A grade, of 14. The 3A class above referred to showed a median

of 17 — or showed a median halfway between the 3A and 4B "norms." Evidently this particular class is above average — or above "standard" — in reading vocabulary, since it scores higher than the average accomplishment of children in the 3A grade.

Another method often used to express the relationship between the scores made by a class and the norm for that class is to find the number of pupils who make a score at or above the norm. Thus, in the class used as an example, 31 out of the 35 children scored at or above the median for their grade — which was 14. Or, to put the matter more clearly, 89 per cent of the children in this sample class score at or above the median for their grade.

The teacher should do more, in the way of a comparison with the norms, than to compare medians, however. She should measure up each child against the whole series of norms. She will then have brought home to her, with new force, the great differences in ability among the children in her class. For the class presented above, the following table might be made to show the relation of the class to the entire series of norms:

1 child	in the class	scores	below	the median for	the 2B grade.
2 children	in the class	score	below	the median for	the 2A grade.
4 children	in the class	score	below	the median for	the 3B grade.
5 children	in the class	score	above	the median for	the 4B grade.
1 child	in the class	scores	above	the median for	the 4A grade.

From this table it can be seen that the scores of this particular 3A class scatter through six half grades. Theoretically, the pupils in the 3A class should make scores between the median for 3B children and that for 4B children; actually, they almost never do. But such a wide distribution as that presented above certainly means that the children differ so widely in

ability to read that they can hardly be taught as a single group. The need of special help for the dull children, and of special opportunity for the bright ones, is strikingly exhibited.

2. TABULATING, AND INTERPRETATION OF TABLES

In the example just cited results from only one class were involved; and the class was very small. Suppose, however, that all the children in a large school from 3B through 4A had been examined. To handle all these results by arranging the papers in order of score would involve an impossible amount of thumbing over of test blanks; and, even after the papers had been so arranged, or after a list had been made from them, any attempt to summarize results would be extremely confusing. Such a list might run as follows:

3B 15, 22, 12, 10, 6, 15, 12, 9, 20, 8, 12, 12, 11, 5, 14, 11, 9, 3, 15, 12, 18, 6, 0, 12, 13, 9, 10, 9, 8, 10, 6, 1, 17, 9, 11, 14, 12, 15, 9, 5, 3, 14, 10, 9, 6, 1, 1, 8, 12, 9, 0, 3, 10, 9, 14, 3, 10, 9, 18, 3, 10, 5, 14, 13, 9, 2, 12, 12, 10, 2, 9, 3, 20, 10, 7, 8, 9, 3, 10, 13, 14, 10, 2, 8, 11, 11, 10

3A 17, 9, 13, 20, 22, 15, 11, 2, 21, 21, 17, 14, 20, 16, 16, 13, 19, 20, 16, 9, 10, 7, 7, 16, 21, 17, 14, 11, 16, 9, 13, 17, 24, 15, 11, 0, 20, 13, 10, 6, 17, 14, 14, 11, 7, 13, 13, 19, 22, 18, 15, 12, 10, 11, 11, 11, 19, 16, 13, 15, 22, 18, 15, 12, 4, 12, 19, 19, 24, 19, 17, 20, 19, 16, 18, 18, 15, 13, 13, 13, 21, 21, 16

4B 27, 20, 15, 28, 20, 23, 27, 29, 21, 18, 24, 27, 24, 21, 16, 28, 16, 21, 24, 30, 21, 25, 20, 25, 21, 21, 21, 28, 20, 19, 12, 7, 27, 22, 30, 29, 25, 18, 29, 25, 18, 18, 19, 16, 20, 24, 27, 19, 22, 26, 30, 30, 25, 22, 19, 25, 22, 19, 25, 21, 18, 21, 24, 27, 24, 15, 15, 23, 20, 12, 23, 27, 27, 19, 22, 20, 26, 26, 22, 12

4A 31, 27, 25, 26, 28, 20, 32, 31, 25, 27, 27, 25, 31, 32, 24, 27, 25, 29, 26, 23, 25, 31, 32, 27, 33, 34, 24, 26, 29, 27, 31, 23, 26, 31, 29, 24, 29, 34, 26, 29, 23, 24, 27, 29, 31, 17, 26, 28, 28, 26, 19, 32, 32, 24, 30, 27, 27, 24, 33, 27, 24, 30, 30, 33, 24, 25, 27, 34, 27, 33, 30, 30, 26

In order to handle these scores the teacher should make up such a table as is shown below (she may either rule such a table, or buy paper so ruled in squares). She should first write in, up one side, the possible scores — in this case from 1 to 34. She should then begin "tabulating," with the 3B papers. The first paper shows a score of 15; so she should make a mark beside the "15" on her table. The next paper shows a score of 22; so she should make a mark in the "22" row — and so on. When she has finished, the record for the 3B pupils will be the first column of marks, as shown on the chart, in the "3B" column. She should then count up the number of marks for each row and write this in to the right of the marks; thus there is one child in the 3B grade who makes a score of 22, there are two who score at 20, and so on. The sum of this column gives the total number of children in the grade. After the record for the 3B is completed the 3A papers should be tabulated, the number of marks in each row counted up, and the total number of cases obtained, in just the same manner. Then the same thing is done for the remaining two grades.

Perhaps there is no single method which the teacher will find more useful than this simple scheme for recording a set of scores. Such a table reveals at a glance facts which she would very likely not discover at all if she had not made the tabulation. The strik-

CHART I

Grades

Score	3B	3A	4B	4A	
34				/// 3	34
33				//// / 4	33
32				//// 5	32
31				//// // 7	31
30			//// 4	//// 5	30
29			/// 3	//// / 6	29
28			/// 3	/// 3	28
27			//// /// 8	//// //// // 12	27
26			/// 3	//// /// 8	26
25			//// // 7	//// / 6	25
24		// 2	//// / 6	//// /// 8	24
23			/// 3	/// 3	23
22	/ /	/// 3	//// / 6		22
21		//// 5	//// //// 9		21
20	// 2	//// 5	//// // 7	/ /	20
19		//// // 7	//// / 6	/ /	19
18	// 2	//// 4	//// 5		18
17	/ /	//// / 6		/ /	17
16		//// /// 8	/// 3		16
15	//// 4	//// / 6	/// 3		15
14	//// / 6	//// 4			14
13	/// 3	//// //// 10			13
12	//// //// 10	/// 3	/// 3		12
11	//// 5	//// // 7			11
10	//// //// // 12	/// 3			10
9	//// //// /// 13	/// 3			9
8	//// 5				8
7	/ /	/// 3	/ /		7
6	//// 4	/ /			6
5	/// 3				5
4		/ /			4
3	//// // 7				3
2	/// 3	/ /			2
1	/// 3				1
0	// 2	/ /			0
No Cases	87	83	80	73	
Median	10	15	22	27	
Norm	10	14	21	26	

ingly good score of one 3B child who makes a score of 22, the scattering of very poor scores in the 3A grade — all such details stand out clearly. So does the relative standing of the different grades. The whole situation, in all its interrelations, is pictured to her.

Finding the median from a distribution table. The table shown in Chart I is usually referred to as a "distribution" table because it shows the arrangement, or "distribution," of marks for a given series of papers. The finding of the median score is much easier if the scores have been tabulated than it is when one has to turn over papers. In the 3B grade there are 87 cases; this means that the middle case is the 44th. To find the middle case, one should simply start at the bottom of the column of numbers showing the number in each row of the table, and add till one reaches the row in which the middle case is located. Thus — the reader should follow on the table — 2 plus 3 makes 5, plus 3 makes 8, plus 7 makes 15, plus 3 makes 18, plus 4 makes 22, plus 1 makes 23, plus 5 makes 28, plus 13 makes 41; since one wishes to locate the 44th case, and since there are 12 cases in the next row, and since there are 41 cases below those in the next row, it is evident that the median — or 44th — case must lie in the "10" row. Consequently the median score is 10. For the 3A distribution, one adds up the numbers through the "14" row; through this row there are 37 cases. Since the median is the 42d case, it is evident that it must lie in the next — or "15" — row. That is, the median score for the 3A class is 15. Medians for the other two distributions are 22 and 27; the reader should reckon these for practice.

For many practical purposes such calculation of the median in terms of whole numbers is all that is necessary; as will be pointed out shortly, test results are only approximate and do not warrant over-exact treatment. However, in comparing groups a refinement of method, permitting calculation to one or more decimal places, often becomes desirable. Such a method is explained, for those who may need it, in the Appendix.

Comparing the results with the norms. On the chart just presented the norms for the different grades have been shown by dotted lines across the distributions for the half grades. The teacher can see at once that all grades, except the 3B, test above the norm, and that the 3B tests at the norm. She can also make certain comparisons as to the standing of children who score at one extreme or the other. Thus, the child in the 3B making the highest score appears above the standard for 4B, while there is a child in the 4B class who scores below the standard for 3B. In general, the "overlapping" of the scores of one grade on the scores of the other grades is clearly shown. A graphical presentation of this sort gives a much better idea of what the grade standing is than can be gained in any other way.

Grouped distributions. One further thing remains to be explained about tabulating, and that is the usual method of "grouping" the scores. The table given above for the four grades is ungrouped; that is, every number from 0 through 34 was given a space — or row — to itself. It would be entirely possible to group this tabulation by twos, threes, fours, or fives. In fact, it would be better to group it in some way, because the table as it stands is a little long to be convenient. If it is grouped by twos, all the 0 and 1 scores would be

put in the same space, all the 2 and 3 scores in the next space, all the 4 and 5 scores in the next, and so on, instead of having one space for each number. The above table would then look as it appears below:

CHART II

SCORES	GRADE 3B	GRADE 3A	GRADE 4B	GRADE 4A
34–35				3
32–33				9
30–31			4	12
28–29			6	9
26–27			11	20
24–25		2	13	14
22–23	1	3	9	3
20–21	2	10	16	1
18–19	2	11	11	1
16–17	1	14	3	1
14–15	10	10	3	
12–13	13	13	3	
10–11	17	10		
8– 9	18	3		
6– 7	5	4	1	
4– 5	3	1		
2– 3	10	1		
0– 1	5	1		
Total No.	87	83	80	73
Median	10	15	22	27

(In this table the marks have been counted and the numbers put in their place.) This table is only half as long as the preceding one and is much easier to handle because of this shorter length. In this particular case

the total number of possible points was 34. But some scales have as many as 200 or 250 points, in which case a teacher of a particular grade is likely to have a range of scores running from 114 to 221. When the numbers cover such a range as this, grouping by twos is not sufficient, as there would still be too many spaces. Such a series of scores should be grouped by fives, having spaces that run as follows: 110–114, 115–119, 120–124, 125–129, 130–134, 135–139, 140–144, 145–149, etc., for the rest of the series. Some test results cover such a wide range of scores that a grouping by tens is necessary, but this is not frequent. The most frequent groupings are by twos and fives; the "ungrouped distribution" is also very common. In general it may be said that one should have from 15 to 25 spaces — or rows — in a distribution table. Thus if the scores range from 45 to 61, the teacher should not group the scores at all, as there are only 16 possible rows in all. If the scores should range from 15 to 95, a grouping by fives would be appropriate, resulting in 19 rows. If the scores should run from 43 to 74, a grouping by twos should be used, thus reducing the number of rows to 17. In fact, the range of scores should be considered and such grouping used as will result in about 20 rows in all.

The finding of the median for a grouped distribution is a little more difficult than is the case with an ungrouped distribution. As before, the teacher should find the middle case — in the instance of the 3B distribution, the 44th case. Counting up the cases as before, one finds that there are 41 cases below the 10–11 interval. In this interval there are 17 cases, among which, if they were arranged in order, the median case would be the 3d, since

there are already 41 cases below this row. Since there are 17 cases in all, the third case up from the bottom will undoubtedly be a 10, so the median may be considered to be 10. If one wishes to be exact, she may go back to the original papers and find out how many of the 17 scores in question *are* 10's and how many are 11's, and reckon the median as before. But this procedure is usually a waste of time if the number of cases being dealt with is small and no great exactness is demanded; the better way is to guess where the median will fall. Thus in the 3A distribution there are 33 cases up to the 14–15 row; in this row there are 10 cases, of which the 8th case is the median case. Since there are 10 in all making scores of 14 or 15, the chances are that the 8th case would be a 15; so the median would be 15. By using a little care one can usually locate the median in a grouped distribution without much trouble. For a more refined method of locating the median the reader is referred to the Appendix.

3. EXPRESSION OF RELATIONSHIPS

The correlation table. The tables — or distributions — that have been presented thus far have been "single-entry" tables; that is, they represented only one series of scores. A correlation table is a "double-entry" table; this means that it presents *two* scores for each child.

PUPIL NO.	FORM A	FORM B	PUPIL NO.	FORM A	FORM B	PUPIL NO.	FORM A	FORM B
1	9	9	10	6	6	19	2	4
2	9	9	11	9	8	20	2	2
3	6	7	12	4	5	21	1	1
4	8	9	13	5	5	22	2	1
5	8	7	14	5	6	23	5	5
6	7	8	15	6	5	24	3	3
7	6	7	16	4	5	25	2	3
8	7	7	17	3	4	26	1	2
9	6	8	18	3	3			

One distribution is arranged horizontally across the top of the page; the other is placed vertically on the side of the page. And the mark representing each child is put so that it will be in both the proper row *and* the proper column. It may be supposed that the two sets of scores above represent the scores made by the same 26 children on two forms of the same test.

The correlation table would appear as shown in Chart III *a*.

CHART III *a*

	Form B									
Form A	1	2	3	4	5	6	7	8	9	
9								1	11	3
8							1		1	2
7							1	1		2
6					1	1	11	1		5
5					11	1				3
4					11					2
3			11	1						3
2	1	1	1	1						4
1	1	1								2
	2	2	3	2	5	2	4	3	3	26

In marking in the cases for this table the teacher should proceed as follows: Since the first child in the list made scores of 9 on Form A and 9 on Form B, she should make a mark in the 9th row and in the 9th column. The second child also makes scores of 9 and 9; for this child she makes a mark in the 9th row and the 9th column. The third child makes a score of 6 on Form A and 7 on Form B; so the teacher makes a mark in the 6th row and the 7th column. A comparison of the remaining pairs of scores with the

correlation table given above will show that the mark for each child is placed so that it may be interpreted as being in both a row and a column — or so that one may tell by looking at the mark what scores the child made on each of the two forms.

The teacher should note that in this table there is a distinct grouping of the scores along a diagonal from the lower left-hand to upper right-hand corner of the

CHART III *b*

	— — — Test 1 — — —								
Test 2	3	4	5	6	7	8	9	10	
14			1			1	1		3
13		1		1		11		1	5
12			1		1		1		3
11	1		1	1	11		1		6
10		1	1	11	1	1			6
9		1		11		1		1	5
8	1		1		1		11		5
7		11		1		1			4
	2	5	5	7	5	6	5	2	37

A comparison of this table with the preceding one should emphasize the complete lack of any trend from the upper right to the lower left-hand corner. As a matter of fact, there is practically no relationship at all between the two sets of scores here presented.

chart. The greater the agreement between two sets of scores, the more distinct this trend. If the relationship is perfect, the marks on the correlation table group clearly along such a diagonal; if there is no relationship

at all, there is complete absence of such grouping. After a teacher has made a "double-entry" table, she should examine it to find if it has any such trend; and she may accept as a working rule that any trend so slight that she cannot see it is not worth considering.

Chart III *b* shows a correlation table in which there is almost complete lack of relationship.

Chart IV illustrates about the degree of relationship that is most common in educational work.

CHART IV

Scores in Mental Ability

Scores in Attainment	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	Total
100–110						1*				1
90–99								2	2	4
80–89					1*		5	1	1	8
70–79					2	2	3	9		16
60–69					1	3	8	4		16
50–59					2	7	7	6		22
40–49		1*	1*	1	2	7	10	8		30
30–39			1	2	8	5	5	4		25
20–29		2	2	2	3	5	3	1	1	19
10–19	1			2	1	2	1			7
0–9	2		1							3
Total	3	3	5	7	20	32	42	35	4	151

The correlation coefficient. The above double-entry tables exhibit the relationship between two sets of scores. Such a relationship is usually summarized in

terms of the "correlation coefficient," which is calculated from such tables. The method of calculation is too complicated to present here. The teacher should understand, however, the significance of coefficients of various amounts. And she should appreciate that these coefficients are simply summaries of facts which appear in detail in such tables as have just been presented. A coefficient of +1. shows perfect relationship; the child who scored highest on the first test scores highest on the other, the child who scores next highest on the first scores next to the highest on the other — and so on down to the child who scores at the bottom on both tests. A zero coefficient — or one of about 0 — indicates that there is no relationship; the child who scores best on one test may score anywhere on the other. A coefficient of −1.00 would indicate a complete inverse relationship; the child who scored highest on one test scored lowest on the other, the child who scored second best on the first test scored next to the lowest on the other — and so on down to the child who scored lowest on the first test and highest on the second. In work with tests, correlations are usually plus; and, roughly, coefficients of .45 or over may be considered to indicate a significant degree of relationship.

The coefficient of the first chart presented was .95; that of the second chart was .21; while the third chart showed a coefficient of .47. The first chart showed a correlation between two forms of the same test; the relationship was very naturally found to be high. The second chart indicates the relationship between scores on an arithmetic test and a rote memory test; there is evidently little or no relationship. In fact, the coefficient .21 may be considered as accidental. A second calculation from similar

data might show a slight negative correlation. The third chart illustrates the relationship, in the second grade of a small city, between the intelligence of the children and their school work as measured by a very comprehensive application of tests in the school subjects. The further significance of this chart will be returned to shortly.

4. APPLICATION OF THESE STATISTICAL METHODS TO PRACTICAL PROBLEMS

The discussion so far has had to do with methods by which the important features of test results may be made evident. There remains need for a brief statement with regard to ways in which these results may be made to indicate desirable changes in school practice. The teacher or superintendent should not only be familiar with the statistical methods described in this chapter; he should also be able to translate the summary statements which he may gather by the use of statistical procedure into terms of any readjustments that may need to be made. He should finally obtain, not a median or a statement of relationship, but an idea of what he should do to solve the problem that confronts him. The remaining section of the chapter is intended to suggest to the teacher or superintendent the various interpretations that may be yielded by the data accumulated.

Comparison of groups. It is something of a truism to say that two groups may differ either in median or in the "scatter" of the scores. But the fact will bear emphasizing. The table below will illustrate certain important points in this connection.

SCORE

	4–5	6–7	8–9	10–11	12–13	14–15	16–17	18–19	20–21	22–23	24–25	26–27	28–29	30–31	32
Class 1 . . .	1		3			5	9	6	6	1	3				1
Class 2 . . .				3	16	14	8	9							

The first of these two groups is the same 3A class presented in the first section of the chapter; the median of this group is 17. The second row shows the scores made by another 3A class in another school; the median of this group is 14. One may express the difference (1) by saying that the median for the first class is 3 points higher than that of the second; or (2) by saying that 31 out of the 35 children in the first class — or 89 per cent — score at or above the median for the second class. Or, it is possible to compare both classes with the standard score as to the distance of the class median from the norm or as to the percentage above that norm. Whatever the method used, the practical implications are essentially the same. Suppose these results were obtained by the supervisor in studying the teaching in the two classes; clearly, the second class is, as a whole, weak in reading vocabulary. And she should advise class work which would increase the reading vocabulary of the class as a whole.

The two classes above presented also differ as regards the scatter of the scores; the total range for the first class is from 4 to 32, while all the scores of the second class fall between 10 and 19. This difference in range would, in itself, suggest very different teaching problems. The first class evidently needs some kind of special instruction for those children who score at

the extremes. This teacher really needs to have one section for the four very dull children, another for the middle group, and perhaps a special section for those scoring at 23 and above. She should certainly make some provision for the four very poor readers and for the one very able pupil; and, should she attempt to include these children in the regular class work, she would find the class most unwieldy and difficult to instruct. The median of this first class is quite high enough; the teaching problem lies with the adaptation of instruction to the extreme cases. In the second group all the children show about the same amount of reading vocabulary. This means that all these children — though there are 16 more of them than in the first class — could be grouped together without any great injury to any one, except for the fact that the class would be too large for such young children. At the most, the teacher would not need more than two sections. But the general standing of the class is low; what this class needs is a raising of the entire group, not a special adaptation to individuals. Should special classes for brilliant and stupid children be contemplated, they should certainly be located in the first school rather than in the second.

It should be evident that the details of the problems presented by each class would not be noticed unless the results were tabulated. In fact, probably not more than one reader out of fifty noticed these prominent differences in the first class when he read over the list of scores as it appeared in the first section. Evidently, also, two classes or other groups will be found to differ primarily in one of two ways, in median or in "scatter."

In the first case the problem has to do with the class as a whole; in the second case the problem is centered around the unusual cases. Of course, a wide range may be associated with a high median or a low median; similarly, either median may be the central point of a distribution showing a narrow range. But, in general, two outstanding recommendations may come from the study of class results. And the supervisor or superintendent may say, "The class median is too low; bring up the work of the class as a whole," or she may say, "The scatter is too great; individualize instruction."

Comparison of individuals. In so far as the educational problem centers about certain individual pupils it will be desirable to have simple ways of summarizing the standing of these pupils on the tests. The most simple and direct method, and the one chiefly recommended by the authors, consists simply in comparing the score of a child with the *grade* norms. For instance, as has already been mentioned, there was one 3B child in Chart I who scored slightly above the median for 4B. Probably the best way to state this child's score is to say that he shows "4B reading vocabulary." The practical implication that this child could do work in the 4B grade is very pertinent. It should also be noticed that a statement in terms of grade norms is the only statement of standing that gives a very definite indication as to the disposition of the individual; surely one of the greatest needs of the American schools at present is for some more flexible system of grade placement.

Instead of grade norms, age norms may be used. So a score on an intelligence test is frequently ex-

pressed as such or such a "mental age." And some workers are developing age norms for tests in the school subjects. The score on a scale for measuring general intelligence may further be converted into an "Intelligence Quotient." This intelligence quotient, or IQ, is simply the mental age divided by the chronological age. If a given child is really 12 years old and gets a mental age of 9 years, his intelligence quotient is 9 years divided by 12, or .75. This figure means that the child's mental development is 75 per cent of what it should be for his age.

As a matter of fact, this percentage form of statement is being used in connection with other types of data as well. Thus if a child's mental age is 12 years and his score on an arithmetic test is equal only to the performance of a 9-year-old child, then one might say that his "achievement quotient" was .75 — or that he was doing in arithmetic work that was only 75 per cent of what, with his ability, he should be doing.

All these methods have their value. But the really important thing is always to make a tabulation of the results from a class or school and then note the position of the scores made by any children in whom one may — or should — be especially interested. So one learns not only that the best reader in the 3B class shows "4B vocabulary"; it is also clear that only a few children in the 3A grade show vocabulary as good; and the exceptional standing of this child is all the more brought out. Similarly, in working with group tests of intelligence it is best to tabulate by grade, and then compare scores with a view to possible grade readjustment, rather than to convert to mental age or IQ. Finally,

if one is studying the relative position of a child on two tests, or his standing on a test as compared with school marks, it is best to make a double-entry table and then examine this table in detail.

The reader should reëxamine Chart IV as an illustration of the way in which such a table should be studied. In this chart there is a conspicuous absence of marks in the upper left corner. Any child who made scores placing him in this position would be very high in school attainment and very low in general ability — and few children show this situation. There are, however, a few children who, by special application and persistence, have raised their school work to a slightly higher level than one would expect from their degree of general ability. These children have been indicated by asterisks. Thus the child who did the best work in attainment showed only average ability; presumably he was possessed of more than usual persistence. There are, however, a very large number of cases scoring in the lower right corner of the chart. All these children show a much higher level of intelligence than of attainment. They are the lazy, bright children who could do better work, and need to be stimulated and encouraged to do better. The children especially in need of attention have been underlined. Thus the brightest child in the class made a score in attainment that placed him in the third from the bottom row. The chief problem of the teacher is with these children — and there are many of them. Those children who are already doing as well as they can or those who, by extra work and application, are doing even better work than their intelligence would warrant, can hardly be expected to improve their work to any great extent; but the child who has ability and is not using it is the one who needs attention and is, incidentally, the child who is pulling down the class average. The teacher should, then, in studying such a chart, locate those children who could do better work than they are doing — and then see to it that they improve.

In first tabulating scores in the form of a chart such as Chart IV, the teacher should note the trend that the marks are taking and when a pupil's scores are such that his mark falls appreciably out of the trend, as in the starred and underlined positions in Chart IV, the child's initials should be written beside the mark to save looking through all the papers again to see which pupils these are that need attention.

Summary. This chapter has attempted to rationalize certain statistical methods for the teacher. The chapter began by pointing out that much could be learned from the record of a class on a test by (1) simply arranging the scores in order. It is urged, however, (2) that always, even with results from a single class, the teacher should make a tabulation. It was also shown (3) that relationships between tests, or between tests and school marks, or between either of these two results and teachers' estimates, could also be tabulated. Finally, it was urged that (4) in comparing classes or in studying individual cases, the relation of the class or the individual to all other results should be carefully studied in detail from the tabulations. The one theme of the chapter has been that the teacher should always get the results before her in tabular form; at least half her troubles in handling and interpreting data will be eliminated if she will do this. It has been the experience of the writers that teachers tend altogether too much to calculations of medians, or other very special forms of statement; and they would urge instead a careful study of the original data.[1]

[1] Suggestions as to further reading will be found in Appendix D.

CHAPTER FOUR

Use and Misuse of Tests

IN previous chapters the writers have tried to show that a test is a valuable and practical instrument for use in the schools. But tests, like any other tool, can easily be misused; and misuse, with consequent erroneous action on that basis, may have most unfortunate results. The present chapter aims to point out certain common errors.

1. Necessity for standard procedure in using tests

As was said in the first chapter, one of the distinctive features which characterize the test method is the careful standardization of procedure. The distinctive service of the tests, in dealing with many of the problems just listed, can be rendered only because conditions are thus rigidly controlled. Further, it should be pointed out that just because a test differs from the usual school examinations in this particular, teachers are particularly likely to forget the necessity for following exactly the directions which accompany the test. So the need of standard method deserves first emphasis.

Directions must be given verbatim as required. If the directions are not thus closely followed, results are sure to be affected. Suppose, for instance, a teacher is giving a test in arithmetic. The work done by the children will be very different according as the teacher tells them to "work rapidly" or to "be sure to make no mistakes." If she mentions that they must be sure to make their work neat, a further factor is introduced. If she

adds that this is a very special occasion and emphasizes to them the notion that not only their own success but the standing of the school is involved, and says that she will give extra drill to those who do poorly, she may so excite some nervous child that his work goes to pieces. The teacher should study the directions carefully, so that not only the verbal directions she gives, but her procedure in passing out the blanks, having the children write their names, and so on, is as it should be. Then she should explain the test to the children, using *exactly* the words required; she should read the directions *verbatim*. This may seem artificial, and there will be the temptation to paraphrase and put statements into one's own words. This temptation must be firmly withstood. It may be that the printed directions are not as good as they might be — though usually the author of a test has a very good reason for each phrase he uses. But whether the directions are good or poor, they must be followed. Otherwise the norms will not apply, since the data on which the norms were based were gathered using the printed directions.

The point cannot be too strongly emphasized. Not only must the teacher avoid any change in the directions; she must not supplement them in any way. Directions are intended to give a certain amount of help to the children — and no more. And the examiner who reads the printed directions, and then consents to answer any questions the children may ask her, is rendering her whole examination largely valueless. Such directions as are sent out with the test materials should be given clearly and slowly, once. No questions from

the class should be permitted. When the directions are over with, the children being tested have been given as much help as the children were given on whose performance the norms are based; and to add further help would be to defeat one of the chief virtues of a standard test.

If the directions for a test are a little long or difficult, the teacher should spend enough time in preparation to make sure that she can read them easily and naturally. Some examinations require not more than five or ten minutes for such preparation; others require hours. But the teacher should be sufficiently the master of the printed directions so that her reading of them will not be mechanical or stumbling — as such reading would give the pupils less help than they are entitled to.

Timing must be accurate and carefully controlled. One more point remains to be emphasized. Practically all tests have a "time limit." That is, the pupils are to be allowed a certain amount of time in which to work. The examiner should be very certain (*a*) that no one begins too soon, (*b*) that exactly the right amount of time is allowed, and (*c*) that every one stops at the same time. If the class as a whole is allowed too long or too short a time, the scores will not be comparable with the norms; if one or two children start sooner than the others, or work after the others stop, they will evidently have a distinct advantage over their classmates. The matter of timing should always be carefully attended to, if the test results are to be of their greatest value. And the teacher should not be disturbed if no one in her class finishes in the time allotted. Most tests are applicable to more than one grade; therefore

the timing has to be arranged so that the children in the highest grade to which the test may be given will have no more than just time to finish. This means that the children in the lower grades will not finish at all; they are not expected to do so. If the children seem disappointed because they did not get through, it is sometimes well to tell them, after collecting the papers, that they were not expected to finish. In any case, the time limit as stated in the directions should be strictly adhered to.

The directions for scoring should be rigidly followed. In some tests the scoring directions are very general; for instance, any kind of mark — underlining, crossing out, circling, etc. — may be accepted as satisfactory. For other tests only one sort of indication is allowed. But, whatever the directions for scoring may be, they should be followed exactly. Sometimes a certain answer is scored right when another is scored wrong, and a teacher cannot see why one answer is not just as good as the other. It may be that both are equally good, though the chances are that the originator of the test had some good reason for his choice. But, anyhow, the scoring directions should be followed; any deviation from the rule makes the norms presented by the originator inapplicable to the data gained by the user. It should go without saying that scoring must be careful and accurate; no child should be misrated because of an error in adding scores. The best way to avoid mistakes, of course, is to have every paper scored twice; but this is frequently impracticable.

General conditions must be satisfactory for testing. It must also be appreciated that the general conditions,

when giving the test, must be satisfactory. A class should not be kept in for a test while the remainder of the school is having recess. No interruptions during the test should be permitted; and careful preparation should be made — plenty of pencils at hand, a sufficient number of blanks ready, and so on — so that there will be no hitch in the procedure. It is usually unwise to give a test just before the children leave school in the afternoon, or at other times when they are excited or restless. The test must never be given under such special situations as would make the results not comparable with the results on which the norms were based.

2. SPECIAL FACTORS INFLUENCING TEST RESULTS

There are many special circumstances that may influence test results. In a discussion of this sort it is impossible even to mention many of these factors. But certain of them are so frequently met, are so constantly to be taken account of, that they necessitate a brief explanation here.

The age-grade situation should always be studied. That is, the teacher should find out whether the children examined averaged old or young for their grade. In some schools there are many retarded children — children who have failed of promotion once, twice, or even more times — and the general policy is against double promotions. The result is that, grade for grade, the children in such a system are older than those in a system which believes in failing as few children as possible and giving extra promotions liberally; the children in the seventh grade in one city may average

as much as eight months older than seventh-grade children in another system. Naturally, the seventh grade in the first system will score the higher on any tests, simply because the children are older and have been in school longer. If a superintendent finds certain schools above standard on the tests, then, he should not allow himself to become elated until after he has investigated the age-grade situation in that school. And, should he find a large amount of retardation, the high scores should be considered possibly no more than a reflection of this fact, and not in the least due to the excellence of the teaching. Clearly, excellence in school work obtained by failing an excessive number of children, and holding the duller ones in the lower grades until they become discouraged and drop out of school, is excellence obtained at a cost, in discouragement and limited education for the duller children, which is most unfortunate. In fact, standard tests have been accused of setting a premium on retardation. If, on the other hand, a class or school tests below standard because the children are younger per grade than those children on whom the norms were based, the situation may be considered creditable rather than otherwise.

No important conclusions regarding the comparative standing of different classes, schools, or school systems should be made, then, without taking into account the age-grade situations involved. Whenever an unusually high or low retardation is found, definite allowance should be made for this factor.

Two methods of dealing with this factor are now being employed. Certain workers are now presenting age norms for all tests, even tests limited in application to only a

> few grades in the school subjects. Another method is to present with the grade norms for a test the median ages of the grades from which the norms were obtained. . . . The whole subject is too technical for discussion in the present volume. It is only possible to warn the users of tests of the importance of this factor, so that some rough account can be taken of its influence in schools showing extreme retardation or acceleration—that is, in schools showing an unusual number of over-age or under-age children.

Test results will vary according to the time of year when the tests are given. It is obvious that a class will do better if tested, in geography, or algebra, or any other school subject, toward the close of the school year than if tested at the beginning of the year, before they have progressed very far in study of the subject. It is also important, in this connection, to take account of the time of year when the norms were obtained. Most tests in the school subjects include, with the norms, a statement of the time of year when the norms were established. If an arithmetic test is given in October and the norms were obtained in May, obviously the norms are not directly applicable to the results; the children tested in October should not be expected to reach this norm. It is best to give such tests at the time of year when the norms were obtained; usually the norms are obtained at that time of year when the test results will be most valuable. However, this is not always possible. If the tests are given at some other time of year, allowance should be made for the difference.

> Such allowance is usually easily made, in a rough way. Suppose, for instance, the norms on a history test were obtained the last of May, and the norm for the seventh

live — and the "pupil material" is of the poorest. It is not to be expected that the children in the latter school will do as well in arithmetic or reading or spelling as those in the same grade in the former school, even though better teachers be assigned to the school in the poorer neighborhood. When, as is more likely to be the case, the inexperienced teachers are assigned to the school in the less desirable neighborhood the standing of that school in the various subjects is certain to be low, and the teachers discouraged. It should then be remembered that the chief factor in the situation is probably not the inexperience of the teachers but the dullness of the pupils. And judgment on the teacher should be accordingly charitable. The point has been already discussed in the chapter on problems and need not be further elaborated here. Group tests of intelligence are particularly useful in making clear any such situation.

Once the intelligence tests have been given, allowance for the above-mentioned factor is easily made. If, for instance, the seventh-grade children in the poorer school show a median score on the group scale of intelligence not better than the sixth grade in the good school, then it should not be expected that the seventh grade in the poor school should do much better work than the sixth grade of the good school on a test in the fundamental operations of arithmetic.

It should, perhaps, be specifically pointed out that allowances of this sort are not to be made, in studying results from tests of intelligence. The intelligence tests are to measure just such differences in ability.

3. PROBLEMS OF INTERPRETATION

After a test has been correctly given and scored, and important factors qualifying the general run of the results, such as those mentioned in the preceding section, considered, there still remains the problem of interpreting the results with reference to practical application. Certain common difficulties deserve careful consideration.

Test results must be used along with, not to the exclusion of, other sources of information. In the first place, use of tests should not be taken to imply complete mistrust of the teacher's judgment. If, for instance, John scores as the best of his class on a group scale of intelligence and if John has, nevertheless, been rated by the teacher as below average in ability, one must not conclude forthwith either that the teacher's judgment is absolutely unreliable or that the test is valueless. The probability is that both teacher's judgment and test score express important facts with regard to the educational problem presented by John. It may be that John is uninterested in his school work and discouraged about it and consequently appears very dull in class; but he may be quick and ready in his dealings with other boys, and have a really extraordinary knowledge of electrical apparatus. The boy is intelligent, but not interested in "book learning." Both facts are needed in order to understand John. And the thing to do under such circumstances is not to reject entirely either the teacher's judgment *or* the test findings, but to understand the discrepancy. It should be appreciated that if the test always agreed with the

judgment of the teachers there would be no use for the test. It is because tests give information which could not be obtained otherwise that they are valuable. However, tests should not be used *in place of* the teacher's judgment; rather they should be used to help the teacher to make her judgments more adequate.

As a matter of fact, both teachers and tests make mistakes. Sometimes a child may be ill on the day a test is given and not able to do himself credit. Occasionally — less frequently than one might suppose — one finds a child who becomes nervous in taking a test in which speed is an important factor, and so fails to do as well as he should. Sometimes the merest accident may upset a child's work; he may break his pencil at a critical moment, for instance. Tests are not infallible, of course, and they should not be expected to be. And they must be used not blindly, but sensibly and intelligently, with due consideration for other sources of information. Once again — tests are not to take the place of other sources of information regarding a child and his work, but to make that information more complete.

In particular, in forming an opinion about the usefulness of a test, one should beware of the "fallacy of the dramatic instance." Thus, in measuring the intelligence of the children of a school by means of a group scale, one may find some striking example where the test is obviously in error. One is then tempted to conclude, forthwith, that the test is wholly unsatisfactory. And, as a matter of fact, a teacher is often heard condemning a test on just such evidence; she knows that John is bright and yet John made a low score — so all tests are untrustworthy! She forgets that the tests gave highly reliable ratings on all the

other children in her class. She forgets also that she herself makes mistakes occasionally; yet she makes no such sweeping condemnation of her own judgment. Argument against tests on the basis of the dramatic instance is very common. One should examine the value of all the measures instead of seizing upon some striking discrepancy, and should aim at a balanced judgment regarding the total efficiency of the test in question.

Tests give not extremely exact but only approximate measures. Teachers making use of tests for the first time are particularly likely to be troubled with the question as to what difference between scores may be considered significant. A teacher is, for instance, dividing her class into "fast" and "slow" sections on the basis of scores made by these children on a group scale of intelligence. She finds that the middle score in her class is 42, and she plans to put the half of her class scoring above 42 in the "fast" section and the half scoring below 42 in the "slow" section. However, Henry, who she thinks belongs in the "fast" section, makes a score slightly below the median score — at 39. And James, whom she has found slow in his school work, scores at 43. The question is as to whether she should follow the test findings or her own judgment in this border-line situation. She should realize that a test gives a measure that is relatively rough. If the test were given a second time to her class, it is quite possible that Henry would score 4 points higher and James 2 points lower. In fact, for most tests, differences of only 2 or 3 points are usually without significance in comparing individuals. In deciding upon border-line cases, therefore, the teacher should follow her own judgment.

In studying this problem of significant differences (as in any work with tests) it is always well to tabulate the results and study the question with the tabulation before one. Suppose that the teacher finds that her class ranges over a total of 40 points on a vocabulary test. Probably a difference between two individuals should be more than 4 points before any confidence should be placed in the reality of this difference. In comparing groups, of course, smaller differences may be important; in comparing the sixth grades in two different schools a difference of 2 points may be well worth attention.

If a teacher wishes to investigate this matter of reliability of results on a test, she should give two different forms of the same tests on the same day to her children and notice any differences in the relative standing of each pupil on the two trials. (Most tests now have two forms, alike in difficulty and in the general ground covered, but made up of different questions.) She will probably be surprised at the differences. Such an experiment will make evident to her the approximate nature of test results, and she will realize that not all differences in score are significant. Very likely the child who scored 2d on the first form will score 5th on the second; and the child who scored 13th on the first form will appear in the 9th place on the second. Children will keep their relative positions only in a very general way. The best quarter of the class will be made up in much the same way according to both forms; but there will be many minor differences in standing within the quarter.

Incidentally, the teacher should notice that the class always does considerably better on a second trial with any test. This is important to keep in mind. In a second trial with a test (either the same form or a different one) it must not be assumed that the norms for this second

form are too low if a class seems to test unusually high. Teachers frequently fail to realize that previous experience on the part of the children with the same or a similar test tends to raise the results somewhat.

Summary. It may be said, then, in gathering together the somewhat diverse comments included in this chapter, that standard tests are very valuable when they are properly handled, but that these same tests may be easily and are frequently misused. In using tests one must be sure, first of all, to follow exactly the directions for giving, timing, and scoring. One must, in the second place, take account of any factors affecting the results as a whole. And one must, finally, interpret the results very carefully and cautiously with due regard to all other sources of information.

PART TWO

TESTS IN THE SCHOOL SUBJECTS

CHAPTER FIVE

Tests in Arithmetic

IT is relatively easy to investigate proficiency in arithmetic. It is relatively easy largely because work in arithmetic involves, as an essential element, an answer — and this answer is either right or wrong. Authorities may differ as to the excellence of English compositions, as to the rules of punctuation, and as to the quality of handwriting. But there is no room for doubt regarding the answer to a problem in arithmetic. And this answer is (in contrast to the questions in a silent reading test) a fundamental and integral part of the work.

As a result the problems of measurement in arithmetic are comparatively simple and straightforward, and methods of measurement have developed farther than in many other subjects. Because of these facts it has seemed best to begin discussion of tests in the school subjects with a consideration of tests in arithmetic. Before beginning consideration of specific tests it is necessary, however, to settle upon some systematic procedure according to which the important facts with regard to each test may be taken up. Such procedure should also be useful as a general scheme or outline which a superintendent or teacher may follow in considering any new tests which may be brought to his attention. The writers have found the following very simple outline distinctly valuable as a basis for study of a test. It has the prime merit of simplicity, and it takes up the various points in the order which would

naturally be adopted by the teacher in becoming familiar with a test or scale.

1. PROCEDURE IN CONSIDERING TESTS

First to be considered in study of a test is, of necessity, its *general nature.* This is largely revealed by looking over the test materials — the blanks on which the children work, the booklet of directions, the score sheet, norms, etc. Such inspection informs one with regard to the general nature of the questions asked in the test, the range of facts covered, and so on. One should get clearly in mind, from such inspection, and perhaps from actual trial of the test on oneself, the type of work demanded of the child. Study of the tests should go farther than this, however. There should be careful consideration of the way in which the questions included in the test were selected, whenever such information is available, and some study of the methods used in building the test. To summarize, then, one should first obtain a clear conception of the general nature of the test and the test materials; such preliminary study is an obvious prerequisite for any further consideration regarding that test's value.

It must not be forgotten, however, that a test may be excellent in its general scope and yield valuable information, and still be impracticable for the schoolman to use. The ease with which the test may be given and scored, and the simplicity and directness of such statistical treatment as may be involved, should be very closely taken into account. A test which is difficult to give and score may involve the teachers in weary hours of clerical work, and result in a diversion of

effort and energy from the routine teaching and in the development of an attitude of antagonism toward tests which it may take years to overcome. The schoolman must never fail, therefore, to consider very carefully the *practicability* of any test he may wish to use. It should be carefully evaluated as regards the convenience and compactness of the blanks, the simplicity and directness of procedure, and the general adaptability in all such respects to use by busy teachers or supervisors, in the average school system.

It is obvious that a superintendent must become familiar with the general nature of a test before he can consider any further points with regard to it, and in looking over the test materials he will form an idea of the practical usableness of the test. But consideration of the test should not stop here; the most important point remains. There is still to be considered, very carefully, the exact *use* to which the test is adapted; and definite study with regard to this point is often omitted. It is not enough to know that a test in arithmetic, for example, is carefully constructed of exercises systematically covering the fundamental processes, that adequate norms are available, that the blank is conveniently arranged and easy to score. The specific service which that test can render in a school must be clearly understood. It will then be appreciated that a "general" test is of only limited value and for purposes of a general survey. And if the test is intended to serve as a point of departure and guide for the teachers in improving the work in arithmetic, then a "diagnostic" test must be employed.

Three questions will be asked, therefore, with regard

to the various tests to be discussed in the following pages: (1) What is the general nature of the test? (2) Is it a practicable testing instrument? (3) What use is it best fitted to serve? These three points, (1) *general nature*, (2) *practicability*, and (3) *use*, will be discussed for each type of test.

2. GENERAL TESTS IN ARITHMETIC

In considering the various tests available in the total field of arithmetic, it is first necessary to classify these tests with regard to their scope. Evidently a test can be very general; the effort can be to obtain a general overview of skill in the arithmetical operations. Such a test should include exercises in addition, subtraction, division, and multiplication, perhaps also in fractions and decimals. And even a test thus inclusive would not cover the entire field of public school arithmetic; there is still ability in problem-solving to be taken account of. A general test covering the entire field of public school work in arithmetic should involve samples of all these various types of arithmetical work.

As a matter of fact, work in arithmetic is too varied to be satisfactorily covered in this way in a single test. The general tests try to cover work only in the fundamental processes. The principal facts with regard to these "general" tests may be briefly presented.

General nature of the tests. These tests may seem to the teacher at first glance to be very little different from an examination which she might make up. They consist simply of blanks on which are printed various exercises in arithmetic which the child is to do. There

are columns of numbers which the child is to add; there are other sums to be multiplied, subtracted, or divided. Perhaps there are examples in the handling of fractions. The test differs from a similar set of examples that the teacher might make up chiefly as regards selection. In the test each example has been included only after elaborate investigation. Each example illustrates some particular element or elements in operations in arithmetic, and the items have been so selected and combined that the test systematically covers the fundamental operations. Further, the difficulty of each example is known. So a good score on the test is significant of a rounded development in ability to handle numbers. The test is also differentiated from the ordinary examination in arithmetic in that directions are very explicit, the rules for scoring specific, and comparison with a large body of norms possible.

Practicability. Almost all of these tests are easy to give and to score. They are, then, highly practicable as instruments for extended use in a school or school system.

Use. These tests indicate the general capacity of the pupils tested in handling, rapidly and accurately, the fundamental number combinations. Such general tests do not diagnose particular difficulties; they cannot be used to inform the teacher that a particular pupil or class is weak in "carrying," for instance. They do serve, however, to indicate the general level of efficiency in the fundamentals of arithmetic. So they are particularly useful at the beginning of the year to investigate the standing of the pupils upon entering a

grade, or at the end of the year to determine the efficiency of the teaching.

The two best known of these "general" tests are the Woody-McCall Mixed Fundamentals and the Courtis Series B. The Woody-McCall test is presented on a single sheet, printed on one side. On this sheet are type problems in addition, subtraction, multiplication, and division of both integers and fractions. The scale gives a rough general measure of efficiency in the operations of arithmetic.

The best known of these tests is, however, the Courtis Series B. This test appears upon a small four-page folder, one page being devoted to examples in each of the four fundamental processes. On the first page there are examples in addition (24 exercises in addition of nine three-place numbers); on the second page there are 24 subtractions of eight- from nine-place numbers; on the third page there are the same number of multiplications (two times four-place numbers); and on the last page there are 24 divisions of four- or five-place numbers by two-place numbers. These items have been selected to include all possible combinations of the four fundamental processes. This scale is particularly useful for supervisory use in the examination of groups. It is evidently diagnostic to a certain extent. That is, separate scores are obtained for each one of the four fundamental processes.[1]

3. DIAGNOSTIC TESTS IN THE FUNDAMENTAL OPERATIONS OF ARITHMETIC

Evidently if a child does well in a general test this is good evidence that that child has mastered all the various elements going to make up skill in number work. But if a child does poorly there still remains the question as to why his work is poor. It may be that

[1] Suggestions regarding the obtaining of sample test materials, and regarding further reading, will be found in the Appendix. Collection and study of sample tests are especially urged.

his difficulty is primarily with addition. His difficulty may be still more specific, however. The child may know his number combinations perfectly but not know how to "carry." In fact, one of the striking results of research in arithmetic has been the demonstration that ability in arithmetic is surprisingly specific. Thus, for most people 6×9 is easier than 9×6, although the two processes would seem, at first inspection, to be practically identical.

Certainly, if a test is to point out the specific difficulties of each child, each one of these various elements must be dealt with separately. If the test is thus arranged, so that analysis is possible, a teacher may discover whether a child fails in addition because he does not know his combinations, because his span of attention is too short, because he does not know how to "carry," or for some other reason. The diagnostic test is designed to make possible such analysis. It is intended to diagnose the particular difficulties of each child. The teacher can then give exactly the help needed, and remedy the difficulty at once. The diagnostic test is thus preëminently the teacher's test. It is intended to assist the teacher in making her instruction more efficient.

General nature of the tests. The general plan of these tests is very simple. There are, of course, separate tests in each one of the fundamental operations. But the analysis goes much farther than this; there is a separate test for each important type of difficulty in each operation. Thus, in a certain diagnostic series there are four separate tests in addition of integers. Samples from each one of these four run as follows:

6	5	7	7493
2	2	5	9016
–	2	4	6487
	0	2	7591
	4	6	6166
	–	0	——
		5	
		1	
		8	
		3	
		2	
		9	
		9	
		–	

The first type involves simply the addition combinations. The second type introduces a short serial addition. The third necessitates a long attention span, while the fourth demands "carrying." It is entirely possible for a child to be able to do examples of the first two types rapidly and efficiently and still be unable to do anything with the last two. The general test does no more than indicate that such a child is poor in addition. The diagnostic test splits up the total ability into its elements, and then tests each one of the elements that go to make up total arithmetical ability.

Practicability. These diagnostic tests necessitate very careful timing but are otherwise easy to give. They are not difficult to score. They are, thus, entirely practicable for a teacher to include among her teaching devices.

Use. The function of such diagnostic tests is, evidently, to point out the particular weaknesses of particular children or classes. Thus a teacher may find that John knows nothing at all about multiplication

although he is very proficient in simple addition combinations; Mary may know the subtraction combinations very well, but still be unable to subtract 98 from 132 because she doesn't know how to "borrow." It is in giving such detailed information as this that the diagnostic test is of value. Class weaknesses may, of course, be investigated with even greater exactness. Once a difficulty has been diagnosed by the tests, it should be remedied by the proper instruction and practice. Thus if Mary cannot "borrow," but is reasonably proficient in the simpler combinations, she should be given instruction and practice in "borrowing" — and should, if necessary, be excused from further drill in mere combinations. The test results should, always, be merely the starting point for corrective instruction.

Perhaps the most useful and inclusive of such series of diagnostic tests has been devised by Monroe. The tests appear on four separate folders. The first two folders present first very simple and then more difficult work with integers. The third folder contains tests in operations with common fractions, and the fourth contains tests in multiplication and division of decimal fractions. In the total series there are 21 tests, each test having 12 to 24 examples.

Another well-known set of diagnostic tests is the Cleveland Survey Series. These tests cover much the same ground, but omit decimal fractions. And all the tests are combined into a single booklet.

4. PRACTICE MATERIALS IN THE FUNDAMENTAL OPERATIONS OF ARITHMETIC

Diagnostic tests such as have just been described indicate to a teacher the exact difficulties under which each child labors in his work in arithmetic. It should be

thoroughly understood, however, that these tests do not serve in actually remedying these conditions. But materials are now available that both assist a teacher in such remedial instruction *and*, almost automatically, keep her in touch with this corrective work.

General nature of the materials. Such practice materials consist essentially of diagnostic tests plus practice exercises specifically dealing with each one of the various types of difficulty covered in the tests. At the beginning of the year the teacher gives the first test covering the work included in the first exercises. She excuses from drill those children who "pass" this first test and has the other children practice each exercise until their work is satisfactory, when they are allowed to go on to the next. The first test is then given again; and the work is continued in this way. The children keep their own records, and present the teacher from time to time with a summary of their improvement.

Practicability. The materials for these practice exercises are, all things considered, reasonable in price. The same cards containing the examples to be worked may be used over and over again; so after the total set is bought further expenses are very light. The tests are an actual time saver to the teacher; after the children have once learned the routine, they give, take, and score their own practice materials with only a general supervision from the teacher.

Use. The paramount virtue of these materials is that they allow the teacher to individualize instruction; that is, each child puts his practice on the exercises where he most needs it. The use of such materials enables a teacher to have an entirely orderly class,

each member of which is working on a different set of examples. Moreover, they release the teacher from the necessity of making up drill exercises, and present drill material of a much more consistent nature than she would give her class. Further, such a scheme gives each child a definite series of objectives; each child knows what progress he is making and knows when he has reached the goal toward which he was working. There is thus no "overlearning" or waste of time on work already mastered; immediately upon completion of each exercise the child takes up the next one. Each child thus progresses through the exercises at the rate best suited to his ability; and the more able children are excused altogether from those portions of the drill that they do not need.

The Courtis and the Studebaker practice materials are the best known. The Courtis set consists of the teacher's manual, students' record and practice pads (of transparent paper), and cabinet of cards upon which the lessons are printed on one side — with the correct answers appearing on the other. All the examples on a single card are of the same type. The child drills on each type until he has brought his work up to standard. In the drill he simply slips the card containing the examples on which he is to practice under the first sheet of paper. The problems printed on the card show through the thin tissue. He then solves the problems, writing down the answers on the paper. When he has finished he turns the card over and slips it back under the paper with his answers on it. The correct answers will appear just over the ones he has written down. He then compares his answers with those on the card, scores his paper and enters his score on his daily record sheet. The lessons cover systematically work in the four operations with whole numbers; there are also three test cards and four special study cards.

The Studebaker materials are very similar in their general nature. Instead of using transparent paper the child slips a sheet of ordinary paper under the practice cards and writes the answers down through holes in the card which appear just below each problem.

5. TESTS IN PROBLEM SOLVING

The discussion thus far has had to do only with work in the fundamental processes. However, skill in the fundamental processes is hardly an end in itself; the child should know how to use these operations in dealing with practical problems; he should have some skill in arithmetical reasoning. In the upper grades, work along these last-mentioned lines is stressed. So tests in arithmetical reasoning, or problem solving, are much needed. Much less attention has been given to tests of this sort than to tests in the fundamentals. However, some excellent tests are available.

General nature of the tests. The test usually consists of a small folder or single sheet upon which are printed problems in arithmetic typical of those studied in school. In fact, such problems are usually selected directly from school arithmetics and are incorporated in the test only after careful consideration of their importance. Special study is also given to the language in which the problems are expressed, that no extraordinary words should appear and involve the child in difficulties of vocabulary as well as of arithmetic. Indeed, the test may be considered primarily a very careful sampling of ordinary school work in arithmetic.

Practicability. The cost of such tests is slight. Giving is easy; there is little difference between giving such a test and giving an ordinary examination. Scoring is

also fairly easy; it is certainly no more difficult than the scoring of a set of ordinary arithmetic papers.

Use. Frequently these tests yield a double score; that is, the child is scored as to the correctness of (1) his arithmetical operations and (2) the principle which he has employed. Other tests give only a single score showing the number of correct answers. But in either case the distinctive information yielded by the problem-solving test is information regarding a child's ability in arithmetical reasoning. Where a test gives a double score, the teacher may obtain a more specific idea of the needs of her class; if the children need practice in reasoning, her procedure in remedying the difficulty will be different from her instruction in case they need simply drill in the fundamentals.

The Monroe Problem Test is typical of the usual test of this sort. This test yields two scores — one for accuracy and the other for "correct principle." By a comparison of these two scores a teacher may be aided in correcting the work of a pupil or class. There are three forms for this Monroe test — one for the third, fourth, and fifth grades, one for the sixth and seventh, and one for the eighth grade.

The Buckingham Scale is very similar. The Starch and Stone Reasoning Tests yield only one score, but are otherwise not unlike the other two above mentioned.

It should be noticed that none of these examinations are diagnostic as to the type of problem involved; they are "general" scales in arithmetical reasoning. Such diagnostic materials are much needed, but no outstanding work along this line has yet been done. Also much needed are practice series in reasoning, analogous to the practice series in the fundamentals.

Summary. As was intimated at the beginning of the chapter, the development of test methods in arithmetic is more complete than in most of the other subjects. First to appear were general tests. Then came the diagnostic tests, aiming at analysis of difficulties and guidance of remedial instruction. Finally, methods of instruction, making use of the test method, have been worked out. Ultimately, the writers believe, such a reconstruction of methods of instruction in all the school subjects will be the final fruit of the testing movement. Already, as will be seen in following chapters, such a reconstruction of the teaching problem has come about in several of the school subjects.

CHAPTER SIX

Tests in the Content Subjects — Geography and History

ABILITY in geography and history may be supposed to involve two elements: (1) knowledge of the facts of geography or history, and (2) capacity for judgment and reasoning with regard to these facts. So there may be two general types of test in this field, "fact" tests and "judgment" or "reasoning" tests. Probably development along one of these two lines would be accompanied by development along the other; but measurement of each type of development would seem desirable.

"Fact" tests are the more easy to construct. Study of the "fact" tests at once reveals, however, the very important difficulty involved in the evaluation of facts in these two subjects. It is obvious that some selection from the enormous number of facts dealt with in these subjects must be made. But the bases for such selection have not been clearly worked out. It is not as yet known which facts in geography or history are the most important for the child to know. It is not known, for instance, which historical facts best equip a child to understand current events and to make him an intelligent citizen. In lieu of selection in terms of such a fundamental objective, test makers have selected facts emphasized in a large number of textbooks as material for tests, upon the assumption that facts included in many texts, written by many writers, would probably be of a fundamental nature. Or, they have simply chosen

facts upon their own judgment. Neither basis is entirely satisfactory.

The points may be made clearer, perhaps, by a consideration of the selection of words for the Ayres Spelling Scale. The words for inclusion in this scale were found by determining the most frequently used words in business letters, newspapers, and compositions. Evidently it is these most frequently used words that it is most important for a child to know how to spell. Selection of words for a spelling test on the basis of one's private judgment regarding the need for learning to spell those words would evidently be a very unsound procedure. To make up such a scale from words most commonly found in spelling books previous to the time of the investigation would have been, in some respects, even more unfortunate; it would have simply perpetuated the earlier notion that children should study the words which were hard to spell, whether those same words were ever used in later life or not. In analogous fashion the value of geographical and historical facts needs to be determined primarily in terms of the usefulness of these facts to the average adult.

1. TESTS IN GEOGRAPHY

General nature of the tests. A considerable variety of tests in geography are now available. These tests vary all the way from series of simple questions such as 'Name the five Great Lakes of North America" to somewhat elaborately arranged test papers showing maps on which the children are to indicate the position of cities or the centralization of industry in a given region. Tests stressing judgment and reasoning rather than fact have also appeared, involving questions such as "Why would you not expect Russia to have as many sailors as England?" Such reasoning tests may be put

in more elaborate form, of which the following will serve as an example (the children are told to "place a cross in the parentheses at the left of the correct answer; if more than one answer seems correct, check the one you think is best"):

Cement-making tends to be a local industry because:
( ) The raw materials needed are found in over half the states.
( ) It is expensive to transport.
( ) It is used everywhere. (*From Witham's Standard Geography Tests.*)

The items to be used in these tests are usually selected from an examination of the curricula of schools, or from an analysis of textbooks, or from a combination of the two methods.

Practicability. For some scales the pupils do not need separate blanks, and in such cases the cost is altogether negligible. Giving is easy. Scoring is simple except in those cases where the answers are written. The tests are, then, practicable, for the most part.

Use. The uses to which these tests may be put may be inferred from their general nature. Most of the tests aim primarily to give a general measure of attainment in geography. And results are more significant than results on ordinary school examinations, chiefly because the questions have been very carefully chosen, their difficulty determined, and extensive norms obtained. Some of the scales offer some diagnosis. So, separate tests may cover North America, Africa, Asia, etc.; and each test may deal separately with location of countries and cities, natural features, industries, and so on. By study of the scores made on the various tests or parts of tests a teacher may discover which points have been left unemphasized and are in need of further dis-

cussion. She may thus save herself from teaching things the pupils already know, and have extra time to put upon instruction along such lines as are unfamiliar.

The Hahn-Lackey scale mentioned in the first chapter may serve as an example of work in this field.

2. TESTS IN HISTORY

General nature of the tests. Again the exact nature of the materials depends upon the particular scale used. The Hahn History Scale presents questions in columns, just as was the case with the Hahn-Lackey Geography Scale. Other history tests appear upon small blanks; usually a scale is made up of several tests devoted either to different periods or to different phases of history. As with the geography scales, the chief problem is the selection of materials, and the chief value of the test is that the questions were selected with extraordinary care. Analysis of textbooks and of curricula is again employed to give the test builder some idea of what, in the opinion of competent judges, might be considered the basic facts of history.

Practicability. None of the history tests with which the writers are familiar are expensive. Those presented upon blanks are easy to give and to take. Scoring, even with tests of this type, is likely to be a little confusing, although tests are now being constructed that permit of a very direct and objective score. In the case of the Hahn Scale, in which the children write out their answers, the scoring is of course more difficult.

Use. The tests in history have the same general values as those in geography. The questions are frequently arranged according to period, so that an ex-

amination of the scores may give a teacher an idea of the particular periods in history which need emphasis. Sometimes a comparison of "fact" and "relationship" questions may tell a teacher something about the type of teaching of which the class is chiefly in need. Scores on these tests hardly tell a teacher the particular weaknesses of particular children, but they do give a reliable measure of general historical information and offer, moreover, a possibility of analysis that may result in further information concerning the class as a whole.

The Hahn History Scale, similar to the Hahn-Lackey Geography Scale, may serve as an example. The chief advantages of this type of scale are, again, a possibility of comparison with the work of other classes and a certainty that the questions asked the class are of importance. Many teachers place a copy of the scale upon the wall of the schoolroom so that the children may know what is expected of them; in this way the scale — like the Ayres Spelling Scale — acts not only as a measuring instrument but as a definite goal to set before the children.

A different type of test is exemplified by the Pressey Tests in Historical Judgment. There are four of these tests — one investigating the ability to read historical material intelligently, a second investigating the knowledge of time-sequence, a third the ability to estimate the importance of events and the fourth the ability to see cause-and-effect relationships. The items for this test were very carefully selected and the questions on the various periods apportioned so that the number of tests on a particular period agreed roughly with the amount of space given to that period in textbooks. The arrangement within each test is by period. The tests are intended, as their name implies, to investigate the extent to which a teacher has been able to develop in the pupils capacity to think

intelligently about historical relationships. Low scores on all four tests would suggest overemphasis on mere facts — particularly if the knowledge of the general sequence of historical facts on the second test should be relatively high. It is also possible to analyze the situation by considering the results by periods.

Summary. As will be realized from this brief chapter, test materials in history and geography are still relatively little developed. The tests are largely general tests; satisfactory diagnostic tests are lacking. And systematically organized materials for the development of knowledge of the various periods or movements have not even been attempted. Scientific devices for bringing about skill in historical thinking and for making the facts of history "function" in relation to everyday problems have hardly been considered. Particularly is there lack of any clear formulation of objectives, with reference to which the value of various test materials might be determined. In the case of spelling, for instance, workers have already determined the 1000 words most commonly used by adults in their usual types of writing, and these words have been incorporated in a scale. Evidently the school should see to it that children learn to spell the 1000 most commonly used words. But just what geographical and historical facts are of most importance in everyday life is not yet known. The whole situation, as regards tests in history and geography, is still ill-defined, and test work is still in its beginnings. But meantime the rough general tests that have been described will be of distinct service to the teacher.

CHAPTER SEVEN

Measurement of Ability in Written English

ONE of the most fundamental tasks of the school is to teach the child to express himself clearly and accurately in written work. And it is one of the most difficult tasks — a task on which a beginning is made in the lower grades, and which the school continues to wrestle with uninterruptedly from then on. Through high school and into college the problem continues; in no other subject is there such continued effort. And one might almost say that in no other subject were the results so discouraging in proportion to the amount of time expended. Under such circumstances effort to express objectives very definitely and to analyze the problem by means of tests would seem particularly desirable.

In discussing tests in arithmetic it was found that three general types of test instrument were available: general tests, aiming at a summary statement of efficiency in handling the fundamentals; diagnostic tests, aiming at analysis of difficulties with a view to remedial instruction; and practice tests, giving materials for the development and control of this remedial instruction. Something of the same situation appears in the field of written English. There are general measures, giving a rough general statement with regard to the quality of written work. There are diagnostic tests for investigating separately ability in grammar, punctuation, spelling, vocabulary, and capable of pointing out special difficulties within each one of these fields. And there are a few series of practice exercises for aiding in

teaching. These various instruments will be taken up in order.

1. MEASURES OF GENERAL MERIT IN ENGLISH COMPOSITION

General nature of the scales. The scales aiming at a summary statement of total ability are very simple in their general nature. They consist simply of a series of type compositions, beginning with a composition which is very poor indeed and progressing to a composition which is very good indeed. These compositions were carefully selected as typical of the various degrees of merit; and their values were determined on the basis of the judgment of experts in the field, an elaborate statistical procedure being employed in order to determine the consensus of opinion among these judges and the equality of the steps from specimen to specimen. The materials consist simply of a sheet on which these sample or type compositions are printed, together with general directions for the use of the scale.

The procedure is also simple. The teacher first has her class write a theme to be rated; the best scales require that this theme be on the same general topic as those themes which have been included in the scale. The teacher then reads over these themes written by her pupils, comparing each with the type compositions in the scale sheet, and she expresses the merit of each pupil's paper in terms of the numerical value assigned to each specimen appearing on the scale. Thus, if she finds that the composition written by one of her children is similar in merit to the composition marked "60" on the scale, she will mark her pupil's composition "60."

The whole procedure is somewhat tedious but very straightforward; the process of scoring is something like matching samples; the teacher merely compares the pupil's paper with each sample on the scale until she finds that sample which is most similar in merit to her pupil's composition. And she then marks her pupil's composition as the sample is marked.

Practicability. It is only necessary to supply each teacher with a scale sheet or scale booklet; there are no special blanks for the pupils. Obtaining the sample compositions from the class is also easy; the teacher should only observe that, for most scales, there are fairly specific directions which she should follow in having the class write the composition, and a definite time limit, with perhaps a special extra period during which the children are to correct any errors. The chief difficulty in using such scales appears in the rating of the compositions. This requires some little care and time on the part of the teacher. However, with a little practice the process of rating becomes hardly more difficult than the usual school method of grading the papers.

Use. The ratings obtained are evidently a synthetic expression of the general merit of the composition rated; general literary value, punctuation, grammar — all these various factors may play a part in determining the rating assigned to a given paper. It is evident also that such measures must be rough; it has been found from actual experiment that different teachers rate the same compositions somewhat differently, and that the same teacher will rate the same composition somewhat differently at different times. However, such variability is less when comparison is made with such type

compositions than it would be if these same papers were graded without such comparison. Such a general measure is evidently useful chiefly in obtaining a rough idea regarding the quality of written work, at the beginning or end of a semester. Or, in large group comparisons (where the large number of cases makes small mistakes on individual papers unimportant, as in educational surveys) one may employ such a method. It should also be pointed out that use of such scales by teachers is of value to them as a method of training, in giving them more definite standards by which to grade written work in the course of the routine classroom teaching.

The Willing Scale will serve as a very simple example of such a measuring instrument. It consists of eight compositions, written by grade school children, on the topic "An Exciting Experience." In using the scale the first thing is to obtain samples from the children. The class is told to write on the general topic, "An Exciting Experience"; and certain special topics under this are suggested, as: "A Storm," "An Accident," "An Errand at Night." The children are given twenty minutes for the writing of the composition, with five minutes more for finishing it, making corrections, and counting the number of words written. The papers are then collected. In rating the compositions the teacher scores them first as to "story value," neglecting errors in grammar, punctuation, capitalization, and spelling. She then goes over the papers again, marking these errors and finding the number of errors per hundred words. Standards are presented both for "story" and "form" values, for Grades 4–8.

A considerable number of such composition scales are now available. The Thorndike Extension of the Hillegas Scale is built on the same general scheme. It is, however, applicable in high school. It also contains a number of

compositions for each of those degrees of merit which are most common. It is thus possible to make somewhat more accurate ratings than would be possible if only a single type composition were given, and to rate compositions on a variety of topics. The Hudelson English Composition Scale is presented in a booklet, together with directions, norms, and (an admirable feature) a set of compositions on which the teacher may practice rating; these samples have been graded by expert judges so that she may check up her ratings by comparison with the values assigned by these experts. The Harvard-Newton scales cover not only narration, but also description, exposition, and argumentation. With each sample composition there is included an analysis of the merits and demerits of that composition. These scales are somewhat limited in their applicability, however, since the compositions were obtained only from the eighth grade. The Lewis scales have the practical merit of including materials for the rating of both business and social correspondence.

The scales so far described investigate ability in written English in general; they do not analyze the situation so that the teacher will know just what to emphasize in her corrective instruction. Evidently a large number of factors coöperate in making up general proficiency in written work. Certain of these factors, having to do with the formal elements in English composition, are of a very specific character and can be easily investigated. Other elements, such as literary merit, coherence, unity, proportion, and mass, are very subtle and difficult to investigate. The composition scales above described emphasize these last elements — but as a matter of fact the ratings are the result of both formal *and* "story" values. The clearly diagnostic tests are now to be considered. These tests,

as will be seen, deal almost wholly with the formal elements.

2 A. DIAGNOSTIC TESTS: ABILITY IN SPELLING

It should be emphasized that ability to spell is of no advantage in itself — is of importance only as a factor in accuracy in written English. Spelling has in the past been treated as an independent subject, having an independent value of its own, in a way which has tended to no little wrong emphasis. A discussion of tests in spelling properly belongs in a subordinate position, as part of a larger discussion of the formal elements in written work. So in this brief handbook no separate chapter is devoted to the subject; instead, spelling has been placed simply as one section of the present chapter.

The Ayres Spelling Scale is one of the most splendid and remarkable products of scientific studies in education, and has operated to bring about a revolution in the teaching of spelling. So the present section will be devoted largely to the consideration of this scale.

General nature of the Ayres Scale. The scale is presented on a single sheet 14″ x 23″. On this sheet are presented the thousand words in the English language which it is most important that a child should learn to spell. These words were selected on the basis of extended investigation as to the words actually used in written work; business letters, newspaper editorials, and compositions were studied with regard to this matter. The scale thus includes only those words which it is most important that a child should learn to spell, since it includes those words which are most frequently

used in written expression. The importance of such selection cannot be overemphasized. It should be obvious that it is much more important that an individual be able to spell "final" or "region" or "employ" than such a word as "crustaceous" which is never written by the average person. The words are arranged on this scale sheet in columns according to difficulty, and the difficulty of each column, for each grade, is indicated.

Practicability. The Ayres Scale is now in such general use that little discussion in regard to it is necessary. Every teacher should have a copy of it; and since the scale sheets are extremely cheap, there is no reason why this should not be the case. In using the scale to test her class the teacher need only select words from the list in which the children of her grade should spell about 50 per cent correct. At least 50 words should be used if possible. These words may then be dictated to the children as in an ordinary spelling lesson, and scored in the same way.

Use. The value of the scale can hardly be overestimated. It informs the teacher with regard to the words deserving emphasis in spelling. And the arrangement according to hardness makes it possible for her to select words for a "quiz" with foreknowledge as to the marks which the class should receive. Thus if she selects her words from the list beginning "catch, black, worm," and she is a second-grade teacher, she should expect her children to average about half of these words correct.

Certain special points with regard to the use of spelling tests remain for brief comment. In the first place it should

be mentioned that, as a result of the centering of attention on these most common words brought about by the Ayres Scale, distinct improvement in the spelling of these words has resulted. So Ayres's norms are probably now somewhat too low. An extension of the scale to include harder words is sometimes desirable. Buckingham's Extension of the scale provides such material. The additional words were selected as common to a number of spellers.

It should also be mentioned that various special methods have been advocated for giving a spelling test. In particular it has been urged that the words should be given in sentences. And various "timed sentence spelling tests" have been devised. Recent research would indicate, however, that such special methods were of comparatively little value, in proportion to the extra time and trouble required. Under most circumstances simple list spelling, using care to pronounce the words distinctly and making them clear by use in a sentence if necessary, appears satisfactory.

2B. DIAGNOSTIC TESTS: KNOWLEDGE OF PUNCTUATION AND GRAMMAR

General nature of the tests. A number of tests in punctuation and grammar are now available. The punctuation tests are all very simple in nature. They consist of sentences devoid of punctuation, printed on a simple blank; these sentences the children are to punctuate. And the score is the number of sentences punctuated correctly, or the number of punctuation marks correctly used. The grammar test may consist of sentences in which the wrong and the correct grammatical form is presented, the pupil being told to cross out the wrong form — as in the sentence, "Was that (he him)?" Or such sentences may be presented, but with a crucial

word left out, the child being told to fill in this word. In another form the children may be presented with incorrect sentences and told to correct them and then give the rule governing the correction. Still another test blank contains lists of sentences, some of which are grammatically correct and some of which are wrong; the pupils are told to underline the wrong sentences.

Practicability. All these various tests are easy to give. Scoring is also fairly easy, though scoring of punctuation is trying on the eyes.

Use. All these tests are evidently of a direct practical value to the classroom teacher. They point out to her the children who are weak in punctuation or grammar. And it is possible for her to go farther, and find out on what particular matters of punctuation or grammar the children need special instruction.

The Starch Punctuation Test consists of a series of sentences to be punctuated, arranged in order of difficulty. The test is valuable as a general measure of ability to punctuate, but it does not diagnose particular difficulties, in use of particular marks. The Briggs English Form Test consists of sentences in which certain marks have been left out, the children being told to add whatever punctuation marks they may think necessary. Provision is made for analytic diagnosis. In order to facilitate diagnosis, the writers have recently devised a diagnostic punctuation test, having four sentences in which semicolons should be used, four sentences in which colons should be used, and so on. It is thus possible, with this test, for a teacher to determine immediately just what type of punctuation troubles a given child. A separate test in capitalizing accompanies this punctuation test.

Among grammar tests may be mentioned Starch's Scale, in which the children are to discriminate the wrong from the right of two words offered in a grammatical construction (as in the example above), and Charters' Tests, in which the child first supplies the correct form for an ungrammatical sentence and may then be asked to give the rule governing the correction. (Either a special diagnostic test giving the pronouns or verbs or a test covering miscellaneous grammatical errors may be had.) The writers have recently devised a grammar test which consists of lines each containing four sentences, one of which is incorrect. The first four lines have to do with the case of pronouns, the second four lines cover errors in the agreement of pronouns, and so on. By combining scores on each group of four the teacher can thus see on what grammatical construction the child is weak. The test has the further advantage of investigating good usage as distinct from formal grammar.

3. OTHER TEST MATERIALS HAVING TO DO WITH WRITTEN ENGLISH

It is obvious that tests in vocabulary have some significance with regard to ability in written expression. Reading vocabulary is much more extensive than writing or speaking vocabulary. Still the vocabulary tests to be mentioned in the chapter on reading ability are of some interest here as giving an indication regarding a child's verbal equipment for purposes of written expression. Particularly important to mention here, however, are series of practice exercises in grammar and punctuation. A number of such practice pads or series of practice exercises are now available, though they are not so closely integrated with a set of tests as to make them proper subjects of discussion in the present brief

volume. "Test and study" exercises in spelling are now to be had; the pupil is given early in the year a large number of words to spell; then, for the remainder of the term, he is allowed to pass over those words he can already spell and to work upon those words which he missed. Such schemes are admirable; but again the discussion of them hardly falls within the scope of this manual.

Summary. There are thus available, as instruments for investigating ability in written English, (1) scales for measuring general merit in English composition, (2) various tests enabling the teacher to investigate, more or less analytically, ability in grammar, punctuation, and spelling, and (3) various practice devices for developing skill in the formal elements of English composition. Evidently the English teacher is not yet supplied with all the instruments that she should have. Tests and exercises having to do with paragraphing, for instance, would seem desirable. Similar materials having to do with the various loose constructions which make up the bulk of the faults of high school and college students are much needed. However, test materials are now available which should be of immense service to the teacher wrestling with the manifold difficulties exhibited by the average American child in written work. And it is unfortunate that such methods have not yet obtained the established place in teaching English which similar methods have gained in arithmetic.

CHAPTER EIGHT

Tests in Reading

BEFORE discussing tests in reading it is first of all necessary to distinguish clearly between oral and "silent" reading. This distinction is of comparatively recent origin. It is only within the last few years that the teaching of reading has meant much more than an attempt to drill children into clearly enunciated and properly modulated "reading out loud" before the class. Ability to read well meant primarily, in the schools of a few years ago, ability to pronounce the words of the reading lesson correctly and to give due regard, in the inflections of the voice, to the punctuation; the extent to which a child understood the meaning of what he read was taken account of only in so far as this affected the expressiveness of his declamation.

It is now tardily realized that ability in reading aloud is for the average adult of very little importance. One may occasionally read a magazine story to a sick friend, or give the family the benefit of an interesting newspaper item at the supper table. But such occasions are rare and of little importance, and for the school to spend much time in preparing the average child for such episodes is absurd. Ninety-nine per cent of the reading done by an adult is "silent reading," or reading to oneself for the thought of the passage. And it is for this type of reading that the school should prepare. Drill in oral reading will not give this preparation; in fact, excessive drill in such declamatory reading may actually interfere with the development of facility in getting the meaning from the printed page.

The present chapter will touch only very briefly upon tests in oral reading; attention will then be turned to the "silent reading" tests. The term "silent reading" is somewhat unfortunate, since it emphasizes a negative aspect of the capacity in question. The important fact with regard to such reading is not that it is silent; the all-important element is the process of thought getting. So it might be better to speak of tests in oral reading and tests in thought getting, or in assimilative reading. This last distinction need not be further gone into here, however; it may best be made clear in discussing the silent reading tests.

1. TESTS IN ORAL READING

From what has just been said it is evident that tests in oral reading are of relatively little importance. Oral reading is of value chiefly in the first grades, as a step in bringing the child to associate meaning with the printed words. So the present discussion will have reference primarily to tests of oral reading for use in the first grades.

General nature of the tests. The most elementary of such tests consist simply of lists of words common to the most widely used primers. The child is shown a list of such words, printed clearly in primer type on a card, and reads each word as the examiner points to it. And the score is determined by the number of words he can pronounce correctly, or by the difficulty of the hardest words he can pronounce. Other tests consist of a series of passages, graded in difficulty from very easy to very hard; the child begins with the easiest passage and reads as far as he is able. The score depends upon rate

of reading and number of mistakes in pronunciation, in omission or substitution of words, and so on.

Practicability. Giving tests of this sort is a tedious and time-consuming process, since each child must be taken individually. Scoring of different types of error in some of the tests is somewhat complicated.

Use. Such tests are evidently of some value to the teacher in the lowest grades, in showing her the number of words which a child can recognize. The type of difficulties appearing in the reading of a passage may also reveal mechanical faults which are preventing the child from making progress in reading. From such oral performance the teacher may infer something with regard to the extent to which the child has obtained the thought of the passage — the extent to which a child understands what he reads. But for this last purpose silent reading tests should, rather, be employed. It is entirely possible for a child to read glibly, and still miss almost entirely the sense of what he is reading.

The Gray Oral Reading Scale is perhaps the best example of an oral reading test. The scale consists of a series of paragraphs, arranged in order of difficulty, to be read aloud by the child. The test must, of course, be given to the children one by one. As the child reads the selections aloud the teacher notices (1) the length of time needed for each paragraph, and (2) the number and type of errors made; six different types of error are taken account of. The final score is a rather complex matter determined from the time and errors, and with reference to the grade in which the child is located.

The scale is of value in checking up the oral reading of children in the first grades. It progresses to very hard passages for use in the upper grades.

The usefulness of such tests is evidently dependent in large measure upon the nature of the method employed in teaching reading. With the growing tendency to teach reading, from the first, with reference to thought getting such tests are evidently becoming of less and less value. As has already been mentioned, a continued emphasis on oral reading may actually interfere with the development of ability in thought getting. So use of oral reading tests in the upper grades is likely to be rather unfortunate than otherwise.

2. GENERAL TESTS IN "SILENT READING ABILITY" OR THOUGHT GETTING

As has already been said, the one all-important purpose in reading is to obtain the thought of the passage read. Evidently the true test of reading ability will investigate this ability to obtain meaning. It is with such tests in thought getting that the remainder of the chapter will be concerned.

General nature of the tests. Tests of "silent reading ability" usually consist of a series of passages about which some sort of questions are asked. The most obvious type of test simply presents a passage and follows this by four or five questions to which the children write out the answers. The following excerpt (from the Thorndike-McCall Scales) will serve as an example:

Read this and then write the answers. Read it again if you need to.

On Monday Dick saw a red fox, a gray squirrel, and a black snake in the woods. The next day he saw a brown rabbit and five brown mice in the field. He killed the fox and all the mice, but let the others live.

9. Did Dick see only one snake?
10. Which animals did Dick see on the second day?
11. Write the name of the day on which the fox was seen.
12. Did Dick see the snake and the rabbit in the same place?

Other tests eliminate the writing by presenting several answers to each question, from which the children select the correct one by underlining it, after the fashion of the vocabulary test to be described shortly. The passages used in the test are usually selected from readers commonly used in the grades to be examined, or from general literature of the type read by children. The tests are scored first of all as to the number of questions answered correctly. This is usually called the "comprehension" score; it indicates the extent to which the child has obtained the meaning of the passage read. A "rate" score is also usually obtained; the most common method is to find the number of words in the passages the child read during the time allowed.

It should, perhaps, be pointed out that measurement of "thought getting" is very difficult. In discussing tests in arithmetic it was remarked that measurement of ability in arithmetic was comparatively easy because every exercise calls for an answer, and the finding of this answer is an essential part in the total process of dealing with the exercise. In reading, the situation is very different. Reading is not normally accompanied by any objective evidence of the degree of comprehension; superficially there is little outward difference between the child who is really understanding what he is reading and the child who is simply staring at a book held upside down. The reading test which follows each paragraph with a question is adding something to the normal reading process. And if a child fails to answer a question it may be because he did not understand the question and not because he did not

understand the passage. This difficulty in obtaining a measure of thought getting should be understood by the teacher, and she should realize that in consequence measures in silent reading are always rather rough. This difficulty is being minimized in the newer tests, however, since the questions are made up only from words included in the passage, and the nature of the questions is worked out with very great care.

Practicability. Giving the various silent reading tests is easy. Scoring is quite simple except for those tests requiring written answers. Tests yielding both "rate" and "comprehension" scores involve rather more labor in scoring, but evidently give more information. There is, then, no practical reason why tests in silent reading should not be extensively used.

Use. The general significance of the results on such tests is obvious. If a class shows poor comprehension, the teacher should consider if she may not be emphasizing oral reading at the expense of thought getting. Whatever the cause for poor comprehension may be, she should emphasize the getting of ideas from the passages read. It may be well to devote the reading period to closely supervised study of the geography or history lesson. The children read a paragraph to themselves, then close their books while the teacher calls on them to summarize what they have read. If rate is low the children may be brought to read rapidly by giving them a large amount of easy, interesting reading matter, which they read to themselves and perhaps take home. It is imperative that there should be interest, as this is the force which will carry them through the material quickly. For a more detailed study of the

situation with reference to remedial measures, diagnostic tests are desirable.

> The general nature of the Thorndike-McCall tests has already been indicated. Perhaps the most widely used are the Monroe tests, consisting of passages each followed by a single question regarding the passage.

It will be noticed that the tests above described are for the most part too difficult for use below the third grade. So far comparatively little has been done toward the development of tests in silent reading for the first two grades. Some admirable beginnings have been made, however. Thus an examination recently issued consists of pictures, beside each of which is some reading matter telling the children to make certain marks upon the pictures. There is, for instance, a picture of a mouse and a squirrel; beside this picture are the words, "Put a ring around the squirrel." Such a test embodies some of the most recent efforts at direct teaching, in the first grade, of reading for meaning. Another test consists of such lines as

tl ut en is de

The children are told to "find the real word and draw a line around it." The test is intended for use in the first grade.

Such tests are new; but they appear to be distinctly useful in following the development of a child's ability to understand and of his reading vocabulary from the very first of his schooling. Still more admirable are efforts at the development of practice exercises, somewhat analogous to the first test above mentioned, for the development of ability to read in young children.

Such methods have been found to expedite astoundingly the development of ability to understand what is read.

3. DIAGNOSTIC TESTS IN "SILENT READING"

Thus far "silent reading ability" has been considered as an entity. But it can hardly be doubted that, just as ability in arithmetic and ability in written English are in reality complexes involving a large number of more or less distinct factors, so also is the ability to comprehend printed matter an exceedingly complex group of capacities. Perhaps the most important single factor is vocabulary; a child — or an adult — understands a passage in proportion as he knows the words in it. Surely, rate of reading and degree of assimilation vary considerably with the interest a child has in the material read. Readiness in silent reading must also be supposed, with younger children at least, to be very considerably affected by the extent to which freedom from "oral reading habits" has been achieved. That is, some children tend, when reading to themselves, to whisper the words, to "read out loud to themselves," and perhaps to follow along the words with the finger. Such habits (usually the result of overtraining in oral reading) tend to slow down a child's reading; true silent reading proceeds much more rapidly than a person can speak. Further, such laborious habits interfere with thought getting.

Curiously enough, relatively little has been done in construction of diagnostic instruments covering these special factors. Recently there has appeared a series of three tests designed to measure separately assimi-

lative reading in history, English literature, and general science. It may surely be supposed that a child's interest in general science, his background of information with regard to this subject, and his general scientific vocabulary might be much more developed than his interest, information, and vocabulary in history. In so far as this was true, such scales should yield diagnostic information of great value. Some effort has been made at tests aiming to determine the presence of oral reading habits. Something has been done toward the development of tests to discover special interests. However, only vocabulary tests have been developed far enough to deserve special mention.

General nature of the tests. All the newer vocabulary tests avoid any writing, and require only of the child that he indicate his answer by underlining or checking a correct definition. So one test (mentioned in the first chapter) contains such materials as the following:

calm (quiet, sleepy, night, restful)
cupola (church, high, schoolhouse, rounded dome)

The children are told to "draw a line under the best definition for each word." A vocabulary test designed for use in Grades 2–4 consists of materials, the general nature of which will be indicated by the following example (the children are told to underline the correct answer):

What does naughty mean? bad pretty fat good
What is a willow? a table a tree an animal a house
What is a pasture? a forest a river a lake a field
What is a monarch? a lady a beast a king a friend
What does dismal mean? reckless naughty sorrowful merry

Such tests evidently permit a child to indicate in a very short time the number of words in the test whose

meaning he understands. It is essential, of course, to include in the test only words which it is important that the child should know. So the best tests are based on extensive research having to do with a determination of the words most extensively used in ordinary reading matter, or the words most common to readers for the grades to which the test is to be applicable. So if the child makes a good score it is evident not only that he knows the meaning of a large number of words, but that he has a practical working vocabulary of the terms that he will most need either in his school work or in later life. It should be added that the words in such tests are arranged in regular order from easy to hard.

Practicability. The giving of these tests is easy and the scoring is simple. So it is entirely possible for any one wishing to investigate the reading vocabulary of a class to obtain and work with such a scale.

Use. The inferences which can be drawn from results of such tests are for the most part obvious. If the score of the class is low, the most natural conclusion (unless special circumstances, such as the presence in the school of a considerable proportion of immigrant children for whom English is not the native tongue, cut across the results) must be that the children have not read extensively enough to acquire a large reading vocabulary. The assignment of a large amount of reading, carrying the children gradually into matter which will introduce them to new words, is desirable. Or there may be a special effort at word drill; the teacher may call the attention of the class to the meaning of the various prefixes and suffixes, and the

relation between words. If the score of the class is scattering, special attention must be given to those children who most lack adequate vocabulary, and methods similar to those described above may be tried especially with those children. Or, it may be discovered that some special factor, such as the hampering effect of oral reading habits, may be the major cause of the difficulty. The special measures previously mentioned may then be employed. Adequate discussion of such remedial measures is outside the scope of the present text.

Perhaps deserving of first mention among vocabulary scales are the recently devised Thorndike vocabulary tests. These are based on very extensive study regarding the frequency of use of words in ordinary reading matter. The Haggerty "Sigma 3" reading examination for use in Grades 5–12 includes a somewhat similar test. The second quotation above is from a very easy vocabulary scale for use in Grades 2–4 devised by the writers.

It was mentioned above that other diagnostic tests are desirable for dealing specifically with other factors conditioning a child's readiness in assimilative reading. A test for determining the presence of oral reading habits would seem particularly needed. The "rate" score yielded by the general tests mentioned in the previous section perhaps gives an indication with regard to this factor. Further work on diagnostic tests in silent reading ability is certainly desirable.

A speed test specifically aimed at such investigation of the presence of oral reading habits, which has been designed by the writers, may be mentioned as suggestive of possible efforts along this line. The test consists of lines of which the following may serve as samples:

They had cold only a few books.
The rabbits read play in the grass.
Once a fairy lived in a wood very.

The children are told to find the extra word and cross it out. The test requires a minimum amount of understanding, so progress is largely determined by speed in reading. And there is evidence that children hampered by oral reading habits make distinctively low scores.

Summary. As has only recently been appreciated, ability in silent reading is of fundamental importance as a prerequisite tool necessary for success in almost all the other school subjects. Only the child with good "silent reading ability" can quickly obtain his geography or history lesson from his textbook in geography or history. Only the child who correctly understands the problems in his arithmetic can solve them correctly. The average American school presents most of its subjects by means of textbooks rather than (as in many foreign countries) by means of oral instruction from the teacher. So ability in assimilative reading is of basic and progressive importance as an immediate need of the school child. As was mentioned at the beginning of this chapter, ability to read understandingly is of the greatest importance for the adult; it is *the* ability which distinguishes him as illiterate or literate. So the schools are under a very great obligation indeed to develop such silent reading ability. An extensive use of silent reading tests in recent years is evidence of the growing appreciation of this fact. The teacher must understand, however, that mere measurement of silent reading ability contributes nothing in itself to the development of such ability. This warning is par-

ticularly necessary since no type of test has been used more indiscriminately or unintelligently. Particularly in use of silent reading tests must the teacher beware of measuring simply for the sake of the measurement. The various factors influencing a child's efficiency in reading are not yet fully understood, and corrective methods are not yet worked out in such clean-cut fashion as is desirable. But this should not prevent the teacher from attempting to remedy any defects which the test may reveal. No one should give such silent reading tests simply because it is the fashion. Before giving such tests there should be careful study of the whole problem; and the test results should here, as always, be considered not an end in themselves but as a point of departure for special teaching aiming at an improvement of the situation.

CHAPTER NINE

Measurement of Handwriting

IN dealing with the subject of handwriting one has to do with the development of a motor habit. The problem of measurement is, then, somewhat different from the problem presented in the previous chapters. The practical requirements in this subject are legibility and speed. Both are important; and speed should not be sacrificed to unnecessary quality. Rather, there should be a definite effort toward rapid writing, with the maintenance of a reasonable legibility.

> In the past, the element of speed has not been sufficiently recognized, and the element of quality has been overstressed. As a result, the children of a generation ago learned to produce a type of handwriting that was almost a kind of drawing. It made no difference how much time was required in the producing of this finished product; and because of the neglect of speed the handwriting of these same children, when they became adults and were forced to write rapidly, deteriorated very badly. Two of the results of the measurement movement have been the discovery of this situation, and the consequent emphasis on speed, in the teaching of writing, above mentioned.

Further, the degree of legibility which is practically desirable has been quite definitely determined. Investigations have been made of the degree of legibility necessary in clerical work in large business concerns and in other occupations. Samples of the handwriting of clerks whose work was satisfactory have been compared with the standard specimens of the Ayres Scale (to be described shortly), and it has been found that a

quality designated by "60" on this scale is entirely satisfactory. It would seem, then, that to train children to write with a quality better than "60" would be a waste of time — time which might better be spent on some other phase of school work.

The measurement movement has thus not only made clear a wrong emphasis in earlier methods in teaching handwriting; it has resulted in admirable practical standards having direct reference to practical requirements in adult life. It is to be hoped that ultimately such standards may be formulated in all the major school subjects. The tests and scales thus serve not merely as a means for measuring the comparative standing of different classes or individual children in a subject; they also define for the teacher the teaching objective. Such a definite objective is of great value to the teacher. If made clear to the child it serves as an aim for his efforts, and so both guides and motivates his work.

The present chapter will first of all (1) present methods for measuring general ability in handwriting. Next, (2) diagnostic instruments will be described. Finally, (3) systematic practice exercises, which may be used in the orderly and systematic development of skill in writing, and which allow for that progress of each child according to his ability which the measurement movement has shown to be always desirable, will be outlined.

1. GENERAL MEASUREMENTS OF SKILL IN HANDWRITING

General nature of the scales. The scales used for the measurement of general quality of handwriting are very simple in their general nature. They consist of large

sheets on which appear samples of handwriting ranging by equal steps from very poor to very good. These specimens were selected from a large number written by school children, and were rated as to quality by competent judges. The scale sheet thus consists simply of a series of typical samples of writing. In measuring the quality of the handwriting of a given child, all that is done is to match the specimen written by the child to those on the scale sheet. The whole procedure is much like matching cloth for dress goods. This process of matching may best be done systematically; the teacher should begin by comparing the specimen with the poorest sample on the scale, and then move the specimen along until she finds that sample which is most like the writing of her pupil in quality. She should then mark on the back of the specimen of her pupil's work the value of that sample. Thus if her child's work is most like the sample marked "50" on the scale, she should mark the specimen "50." It is generally wise to rate samples first by beginning at the bottom of the scale and working up, and then by beginning at the top of the scale and working down.

The above simple procedure is used in measuring quality (the general method is thus similar to the method used in measuring the merit of English compositions). A speed score is also usually obtained; this consists simply of the number of letters written per minute.

Practicability. Every teacher should have such a scale. The rating of samples from a class is a somewhat tedious procedure, but the work comes to be done easily and quickly with practice. Evidently such rating is

more or less inaccurate; different teachers differ somewhat in the ratings they assign, and the same teacher may frequently rate a specimen differently if she rates it a second time. Some preliminary practice is needed to obtain skill in this work; and samples which have been rated by experts may be obtained, with which the teacher may practice. Care is of course necessary in obtaining samples from the class; evidently the writing of the class will differ according as the teacher may direct the class to "write carefully" or to "write as rapidly as you can." The usual method is to obtain the specimens without telling the class that a test is being made, using carefully worded instructions stressing neither quality nor speed. To sum up, then: measurement of handwriting requires exact procedure in obtaining the samples and some little close application in rating them. But the time and effort required are not great, and no one should hesitate to make use of such scales on account of the labor involved.

Use. Such scales evidently yield a general measure of quality and legibility and a statement of the speed at which this quality is obtained. The value of such results in investigating the efficiency of the training in handwriting is obvious. General remedial measures may naturally be suggested by the results. If quality is poor, effort must be made to improve quality; and detailed information with regard to defects may be obtained with one of the diagnostic instruments shortly to be described. If speed is lacking, it is evident that there has been too much effort at "copper-plate" writing. The teacher should then realize the practical necessity for rapid writing. And she may give timed

exercises, and so force the children to write up to a certain rate — or otherwise develop greater facility.

Typical of these instruments for investigating general efficiency in handwriting is the Gettysburg edition of the Ayres scale. The teacher should have the children copy the first three sentences of the "Gettysburg Address" until they are thoroughly familiar with the wording. This is a preliminary exercise. The children are then asked to copy these three sentences again, all beginning together at a signal from the teacher; she stops their writing in exactly 2 minutes. Both speed and quality are taken account of. The number of words per minute is easily found by means of a chart giving a cumulative count of the number of letters through each word. The rating for quality is found as previously described.

2. DIAGNOSTIC MEASURES IN HANDWRITING

The scales so far described yield only a general statement with regard to quality; they do not show specific weaknesses, nor do they offer specific suggestions with regard to remedial instruction. For such analytical study further instruments are necessary.

General nature of the methods. Various devices may be used for such analysis. Thus separate scales have been devised for measuring uniformity of slant, uniformity of alignment, quality of line, letter formation, and spacing. By comparing a given child's handwriting with such varied materials, it is possible to analyze to a very considerable degree the specific type of weakness which that child shows.

The analysis may be approached in a different way, however. So score cards have been worked out. Such cards list separately the various qualities which may be taken account of in judging handwriting, and provide for the assignment of different degrees of merit in each

one of these categories to the particular specimen being judged. Thus the writing of a particular child may be first rated as to heaviness, then as to slant, then as to size, and so on. If the size is very excellent and uniform, a grade of 7 may be assigned to that specimen; if less good, a rating of 6, and so on down to 1 as the poorest rating. By such a scheme each specimen is judged as to its standing in these various respects; and the sum of the ratings assigned makes up a total score or statement of total merit.

Practicability. As will be obvious, the use of these diagnostic instruments requires much time and care. The average busy teacher will, then, hardly find it possible to study an entire class after such a fashion. These special methods are very useful, however, in dealing with special cases showing unusually poor handwriting. The materials are of course readily available for any teacher who feels need of such special methods for studying any children whose handwriting shows lack of improvement.

Use. As has been suggested, because of the great care and amount of time necessary in using these diagnostic instruments, they are hardly practicable for very extensive use. They find their chief value in efforts at the analysis of difficulties presented by very poor writers. By means of such methods it may be determined that the major fault of a child is in slant or alignment. The child's attention may then be centered upon this one feature, and special exercises may be given to remedy this particular defect. Successive applications of such a diagnostic instrument will permit the teacher to follow any improvement that may

appear, and will demonstrate to the child any such improvement and so give him encouragement.

Usually it is best to use the general scales described in the previous section as a method for preliminary sorting out of those cases which need special attention and intensive study by means of a diagnostic instrument. So the Ayres Gettysburg scale may be given to a class, and the general standing of the class and distribution of ratings obtained. This distribution may show two children who are very poor in the quality of their writing. Study of the work of these children by means of the Gray score card or the Freeman diagnostic scales reveals particular defects. The teacher then points out the defects to the children and at each writing period gives them special attention. She further, after two weeks, obtains new samples from these children which she rates by means of her diagnostic scale. The general procedure desirable under such circumstances is obvious; the details a teacher can work out for herself.

The two best-known instruments for diagnosis are the Freeman scale, consisting of samples of three different qualities as regards alignment, letter formation, and so on, and the Gray score card, which permits rating of a specimen as regards nine different features: heaviness, size, slant, alignment, spacing of lines, spacing of words, spacing of letters, neatness, and formation of letters; and under each one of these main heads more specific features are listed. It is suggested that a teacher will find it desirable to obtain sample copies of these two instruments, even though she may not find it possible to use them; they will present her with an analysis of the factors important in handwriting and make clear to her the problems in the teaching of the subject.

3. SYSTEMATIC PRACTICE EXERCISES IN HANDWRITING

It has already been emphasized that the finest fruit of the measurement movement appears in a reconstructed method of teaching which permits each child to advance in proportion to his ability, which permits the teacher, nevertheless, to keep in close touch with the progress of each child, which allows each child to follow his own development and presents him with a clearly defined objective toward which he works. Such schemes have been worked out in the teaching of handwriting. A detailed discussion of such methods is outside the scope of the present text. It need be said here only that, as mentioned at the beginning of the chapter, practical requirements in handwriting have now been quite definitely determined. It is now realized that the schools of a few years ago aimed at a quality of handwriting much above practical requirements, and a quality which could not be maintained under stress of the speed required in everyday writing. It is also realized that to train children to produce a quality of handwriting much above practical requirements, or a quality which cannot be maintained in everyday life, is a waste of effort and a form of overtraining which is undesirable and may possibly be even injurious. Further — as has already been mentioned — the practical requirements of the business world have been definitely stated in terms of the measuring instruments now available and, as with all the standard scales and tests, highly reliable norms for the various grades have been established.

All this information is evidently invaluable as the

basis for proper understanding of the practical problems involved in the teaching of handwriting and the development which may be expected of children in a given grade. It has further been found possible to put such measuring instruments as have been described in the hands of the school children and have them grade their own work and construe their work with reference to these various standards. Such self-guidance on the part of the child evidently relieves the teacher of much routine work which she would otherwise have to perform. It frees the teacher so that she can give special attention to those children needing such attention. Further, there are thus presented to each child definite objectives which he can understand and which supply a motivation that could hardly otherwise be obtained. And as the child not only practices but scores the product of his practice, he thus is enabled to follow his progress and appreciate for himself any gains which he may make.

As has been emphasized again and again, the bane of the measurement movement has been the tendency to test merely for the sake of the testing. And it has been constantly urged that the teacher should always understand that measurement is not an end in itself, but rather a starting point for instruction aiming to remedy any difficulties revealed by the measurement. In most of the school subjects methods of instruction have not as yet been sufficiently integrated with the measurement movement to bring about explicit formulation of methods with reference to the results of testing. So in most cases it is possible only to indicate to the teacher in a general way the application of any test

findings to her teaching problems. Wherever possible, the teacher should aim to obtain for her use any schemes which present to her in expressly formulated fashion such integration of tests and teaching method.

The writers have particularly in mind, in connection with the above brief discussion, the Courtis Standard Practice Tests in Handwriting. The materials include a teacher's manual, a pupil's daily lesson book, a pupil's daily record card and graph blank, and class record sheets. Use of the series begins with a research test to show the initial standing of each child; on the basis of this first test those children are excused from drill who make records up to standard. The other children start with the first practice exercise and work on each exercise until they have attained the "standard" on it as shown by the daily test; as soon as a child reaches "standard" on one exercise he proceeds to the next. Each child is trained to score his daily test and to keep a record of his progress. And at the end of the semester a second research test is given to measure the progress which has been made.

The advantages of such a method are obvious, and the methods and materials are well worth study even by teachers not planning use of this particular practice series in their classes. Such study will be found highly suggestive and instructive with reference to methods both in handwriting and in other drill subjects.

CHAPTER TEN

Tests in the High School Subjects

THE general situation in high school differs fundamentally from the situation in the grammar school, in that the children represent a selected group — since only the brightest children survive long enough in their school career to reach high school; and the pupils taking a given subject are still further a selected group. Moreover, the children themselves have changed since they are no longer children, but adolescents; and the immediate aims of the high school are not identical with those of the grammar school, however similar the ultimate aims may be. It is, then, to be expected that tests for high school should differ somewhat from those applicable to the grades. The subject matter is certainly different, and the form is very likely to be somewhat changed. The reader will doubtless notice, for instance, that most of these tests require writing — a condition which was severely criticized in the earlier chapters. The tests would probably be easier for the pupils to take, and certainly easier for the teacher to score, if the writing were eliminated; but the demand for writing does not invalidate the results to any dangerous extent, since all the pupils in high school write with a fair rapidity. Also, the examiner may take for granted a certain amount of intelligent understanding of directions and the general problem of the tests, for only the more intelligent children get into high school. It will be seen that the tests are very much less special in form and much more like an ordinary class exercise than is possible with work with younger children.

The tests that are of value to the high school teacher are chiefly in the following five fields. There is (1) still the problem of measurement in English. (2) Tests in algebra, of various types, are available, and are of distinct practical value. (3) Geometry tests have appeared. (4) A variety of Latin tests have recently been issued. Finally, (5) tests in the Romance languages are to be presented. The tests in these various subjects will be taken up in order.

1. TESTS IN ENGLISH

The various tests and scales for investigating work in written English mentioned in a previous chapter are almost all applicable to high school; in fact, many of them are best suited for work in these grades. And they are much needed in high school work; high school students write compositions that need to be rated, and make errors in grammar, punctuation, and spelling that should be analyzed and corrected. Training in written English in high school is very frequently desultory and lacking in clearly formulated objectives. Use of tests will present the situation in clean-cut fashion and will make concrete the objectives to be reached. So the teacher of high school English may very profitably read the chapter on written English with special reference to the problems of high school work in composition.

Little test material has yet appeared dealing with the teaching of literature. However, beginnings have been made in this direction, and there will probably soon be practicable instruments for (*a*) investigating a pupil's knowledge of the facts regarding various literary

masterpieces, and (*b*) finding out something with regard to a pupil's ability to discriminate literary merit.

2. TESTS IN ALGEBRA

General nature of the tests. The various tests in algebra now available look, superficially, much alike. They all consist of blanks on which certain exercises in algebra are printed. They differ from a series of exercises that might be printed by some teacher in that the materials have been very carefully selected, either as common to many standard texts or as involving the most fundamental elements in algebra. This last type of examination is not merely a measure of the classroom teaching, but is of value to the teacher as a formulation of objectives in the teaching. By means of such an examination especially the teaching of algebra may be expected to be improved and redirected.

Practicability. The administration of these tests is very easy, as the directions appear on the same page with the examples; the scoring is not difficult. There is thus little excuse on the part of the teacher of algebra for neglecting to make use of test materials.

Use. The best algebra tests are all diagnostic to a greater or less degree; that is, they will inform a teacher with regard to the specific types of problems in which her pupils are weak. The tests will also serve, through a use of the total score, as a general measure of the efficiency of the teaching. The different tests differ considerably in their scope and, consequently, in the significance of their results.

Space is not available in the present brief manual for a description of these scales. The Rugg-Clark scale will

serve as an example of a test based on the consensus of textbook practice; the Illinois (Monroe) test attempts investigation only of the distinctive processes involved in algebra. It is suggested that algebra teachers should obtain samples of both these scales.

3. TESTS IN GEOMETRY

General nature of the tests. The tests in geometry present theorems or parts of theorems that are to be solved by the student. In practically all cases the theorems are unfamiliar to the student. It is then expected that the pupils should have gained from their study of geometry an ability to deal constructively with new problems. In some cases only a general measure is obtained; other scales attempt an analysis of the situation.

Practicability. The directions for giving are usually printed on the blank, so that all the teacher has to do is to pass out the blanks and collect them at the end of the time limit. The scoring of such papers is a little more difficult than is the case with some tests, but is surely no more complicated than the correcting of a similar number of school examinations in geometry; and the scores when obtained tell more about a pupil's ability to demonstrate theorems than can usually be gained from the reading of examination papers.

Use. The general tests in this subject give a measure of the ability of the pupils to solve unfamiliar propositions. By selecting theorems that are unknown to the pupils the possibility of a mere memory production is usually avoided and one is able to get back of the memorizing of proofs to the ability to handle the basic

facts of geometry in a logical manner. Any analysis that may be presented by the tests will be found useful.

The Minnick, Thurstone, and Starch Geometry tests may be mentioned as examples of work in this field.

4. TESTS IN LATIN

There are two types of test for use in studying the products of Latin teaching. These two types may be referred to as tests of the direct outcomes of the study of the language — that is, knowledge of Latin words and ability to read Latin — and as tests of the indirect outcomes of the study, such as the influence on the pupil's knowledge of English. Both types of test have appeared, though the work in measuring the more direct outcomes is older and more developed. Both types of test are of value to the teacher of Latin. The test of Latin vocabulary or of Latin translation is of use primarily for the teacher to determine the extent to which her pupils have learned the Latin grammar that she has been teaching them. But the teacher of Latin hardly expects her students ever to acquire the ability to read Latin with ease. Her best efforts should, perhaps, rather be expended in teaching the child knowledge of Latin words and constructions so that he may be able to understand his own language better. That is, the most valuable results of the study of Latin are not the direct results. They are not the ability to translate or to speak the Latin language. The most important results are those that come indirectly. Study of Latin should make easier the study of modern languages. It should, perhaps, make clearer the life

and contributions to civilization of the classical peoples. Such study may, possibly, give a certain general training in precision of thought and close application. Especially, such study should give a pupil a better understanding of the English language. Many teachers of Latin do not sufficiently emphasize these indirect values. It should be appreciated that the study of Latin will have little effect upon the pupil's mastery of English, for instance, unless specific attempts to relate the two languages are made. So, tests attempting to get at these indirect values of Latin should be of great value; they should measure the extent to which such indirect values have been attained, and should make clear to the teacher the need for a consideration of these objectives. The Latin teacher should, then, not only measure the Latin vocabulary of her pupils and their ability to translate Latin. She should try to find out also the influence of this teaching of Latin upon work in other subjects, particularly English.

General nature of the tests. The best tests for measuring the direct outcome of the study of the language usually deal with either vocabulary or translation. The items for such tests are usually selected from material common to first-year Latin books or to Cæsar, Cicero, and Virgil. The vocabulary tests may be constructed from a list of the words most common to these standard texts, and based on careful counts of the frequency with which each word occurs in such material with which the pupil deals. The tests aiming at investigation of indirect values in Latin are of a very different nature. They require the giving of English words which are Latin derivates, or the spelling of such words,

or their definition. The effect of the study of Latin on knowledge of English grammar is also being studied, and the transfer to the study of Romance languages (though work in these last two fields is still in its beginnings). Influence on English is most important. It will surely be granted that the study of Latin should help the pupil in defining English words or in spelling them, or in understanding English words of Latin derivation. Such tests, it will be seen, attempt to measure what might be called the less tangible results of the teaching of the language.

Practicability. Directions for taking these tests usually appear upon the blanks themselves; the tests are then easy to administer. Practically all the tests call for writing on the part of the children. This means that the scoring of the test is, of course, a little slower than it would be if writing were eliminated. But the amount of information gained by the teacher from the use of such tests is so great that such slight inconveniences of scoring are of no importance.

Use. Those tests designed to measure the direct results of the teaching of Latin obviously give an indication of the extent to which the pupils have learned those words and constructions with which they should be familiar. It should be noticed that the words used in the vocabulary test constitute a selected, and not a general, vocabulary. It would be hardly possible to estimate from such tests the total Latin vocabulary or the total ability for translation of any given child; the scope of the test is limited to those elements of Latin vocabulary and structure that are most common. The tests give very efficient measures of the extent to

which the children have profited by what they have studied. The other type of test, measuring the indirect results of teaching Latin, should serve to indicate to the teacher the extent to which she has succeeded in making the language of real use to her students, as by helping them to understand their own language. If her pupils fail to see the relationship between the two, — if they cannot, for instance, give English derivatives from Latin words, or if they cannot recognize misspellings of words with Latin roots, — it would appear that the Latin they have learned is not functioning as it should in its influence on English. If a teacher finds such a condition in her class she should certainly make attempts, especially if the scoring of such pupils in vocabulary and translation is good, to emphasize relationships between English and Latin so that the knowledge of the classical language already gained by her students will become a real factor in helping them in their work in English.

The Henmon Latin tests and the Kansas Latin derivative tests will serve as examples of tests measuring respectively direct and indirect products of the teaching of Latin; the two series are well worth very careful study by Latin teachers.

5. TESTS IN THE MODERN LANGUAGES

In learning a modern language a pupil is expected to achieve the ability to read ordinary literature in that language. To do this he must attain a vocabulary covering the basic words. Moreover, in order to understand the language, he must be acquainted with the grammatical constructions and sentence structure in-

volved. Thus it would seem that in building modern-language tests one should include measures of vocabulary, grammatical construction, and silent reading or translation. Tests of all three types are now available.

General nature of the tests. Tests in the modern languages usually have two parts; sometimes there is a third, but not frequently. The two most common types of a modern-language test are devoted to measurement of vocabulary and measurement of ability in silent reading or in translation. A third part devoted to grammatical construction is sometimes found. In selecting the words to be included in the vocabulary test the originator usually analyzes beginners' books or general literature in the language involved, and determines those words which appear to be the most common. He then employs these fundamental words in the test which he is constructing. Usually the pupils write out the translation of the words in the vocabulary test. The test in silent reading may be either a brief selection of easy reading material with questions asking about what has been read, or it may take the form of sentences to be translated. Any test of grammatical construction attempts, of course, to measure knowledge of those constructions which are felt to be the most fundamental. Thus in the selection of material for tests in the modern languages an attempt is made to select those words, sentences, or constructions that it is most necessary for the student to know.

Practicability. The material usually consists of one or two sheets of paper or of a small booklet. The giving of such tests is exceedingly simple. In fact, practically all tests for use in high school are easy of ad-

ministration, because directions to the pupils can be printed on the blank, where the pupils can read them for themselves and may return to them whenever they need further information. The teacher therefore acts only as a general supervisor, maintaining order and seeing that the proper time is allowed. The scoring of such papers is not difficult. Of course any handwriting involves the teacher in some deciphering of illegible penmanship. But the demands made upon the teacher by such a test are certainly no greater than those made upon her by the usual written lesson.

Use. Such tests, being based upon the fundamental words of the language, give a teacher a general measure of the knowledge of her pupils of those basic words. She obtains, also, a measure of the ability of the pupils to translate simple sentences. The teacher may feel sure in using such tests that she is not confronting the pupil with unusual or unfamiliar material, for the words have been so rigidly selected that every student should know at least some of them. If the words of the test appear to be unfamiliar to a pupil, it is a fairly good indication that he is not very well acquainted with the fundamental facts of the language. The tests yield, it should be understood, largely general measures —make little attempt at diagnosis. Such measures of a diagnostic nature may be developed before long, as workers come to understand more thoroughly the abilities which go into the learning of a language.

The Henmon and the Handschin French tests will serve as examples of work in this field.

The interest in tests began with measurement in the grammar school subjects. And while the movement

soon spread to high school, the tests for use in high school are not yet, for the most part, so well developed as those for use in the grades. Still, tests should be of great use to the high school teacher; and use of tests with the more mature high school pupils is easier than is such work in the grades. It is distinctly unfortunate that most high school teachers are not as yet interested in test work. Within a few years testing will undoubtedly be as common a practice in high school as it is now in the grammar school.

PART THREE

TESTS OF MENTAL ABILITY

CHAPTER ELEVEN

The Measurement of General Mental Ability

1. THE GENERAL NATURE OF THE TESTS

So far the discussion has been concerned with tests in the various school subjects. And the effort has been to involve in the subject matter of these tests as much as possible of the actual content of these subjects. The aim has been to make the tests measures of the *products* of teaching. Now, a very different type of problem has to be dealt with. In the tests to be discussed in the present chapter the effort has been, so far as possible, to *keep out* of the test any elements which might be affected by schooling. Tests of mental ability should measure one's capacity to do school work, and should be uninfluenced by the extent to which one has or has not had educational opportunity to date. This distinction should be kept clearly in mind. The test of ability tries to get back of the results of education to original capacity. In building tests of general mental ability, the great problem is thus to include only facts and problems which are not dealt with in the school. Perhaps one might put the situation crudely by saying that the tests of general mental ability investigate more one's "common sense" as distinct from one's "book learning." And always the distinctive contribution of the tests of mental ability is made when these tests show an ability different from a pupil's educational status. The suggestion is, then, that some special effort should be made to influence such a child to do school work in accordance with his ability.

The essential nature of tests of general ability is very clearly revealed by study of the practical situation which brought about the formulation of the first successful scale for measuring general ability — the well-known Binet-Simon Scale. In 1904 provision was made in France for the organizing of classes for subnormal children in the Paris public schools. Then the puzzling question arose as to how children would be selected for these classes. It was intended that membership should be restricted to those who were actually subnormal in ability. Children who were backward in school because of frequent absence, because of illness, or on account of some other special handicap, should not be included; such children can, with a little special help, be brought to the point where they can do the regular school work. The special classes were intended only for those who were backward in school and could *not* be returned to the regular classes, because their backwardness was irremediable and was due to mental defect. And the question was as to how these defectives might be distinguished from those who were backward in school merely as the result of some special handicap.

At this juncture the advice of Binet was asked, and his discussion beautifully reveals the concept back of such tests. In selecting children for these special classes Binet urges that three types of examination be given. There should be a pedagogical examination; this should reveal the extent of the child's backwardness in school. There should also be a medical examination; this should reveal any evidences of gross physical defects, and eliminate the possibility that some special physical handicap, such as partial deafness, may be causing the backwardness in school. Finally, there should be "psychological" examination. This examination should be especially intended to distinguish those who are bright but backward in their school work (that is, those who are capable of doing good school work but are retarded for such a cause as irregular attendance) from those who are backward because of mental subnormality.

And Binet's notion is, after all, simple enough. These children who are bright, but backward because of some special reason, will make a poor record on the pedagogical examination. But though they may have lacked educational opportunity they have nevertheless had the opportunity, which every child does have, of learning common-sense facts about the environment from playmates, parents, and others with whom every child comes in contact. It was Binet's aim to test this "common sense." The younger children are asked such simple questions as, "What is a chair?" "Which is your right hand?" "Is this morning or afternoon?" An older child is asked, "What is the thing for you to do if a playmate hits you without meaning to do it?" "In what way are wood and coal alike?" (The writers are quoting from the Stanford Revision of the Binet Scale.) Farther up on the scale a child may be asked what is foolish in such a statement as, "Yesterday the police found the body of a girl cut into eighteen pieces. They believe that she killed herself."

It is obvious that ability to deal with such questions as these would hardly be affected very much by schooling; every child should pick up such information as is involved in these questions. Nevertheless, the questions require a type of ability quite as complex as does the regular school work of children of these ages. The child who is backward in school, but who is nevertheless able to answer such questions as well as the average child of his age, may be supposed to be backward because of some special circumstance, not because of mental defect. And we have an instrument for dealing with the practical problem above mentioned; such tests will serve to differentiate those who belong permanently in a class for defectives from those who merely need special help for a short period.

Such tests, then, are useful in school because they measure ability to do school work as distinct from the extent to which a child has profited by his schooling.

The reader should have this distinction clearly in mind, and no teacher should feel herself responsible if one of her children makes a very low score on such a test. Rather she should feel relieved, since this low score may be taken to demonstrate that the poor school work such a child is doing is the result of lack of capacity on the part of the child, and not of poor instruction on the part of the teacher. In fact, almost all the evidence goes to show that a low score on a test of mental ability means a poor ability which has been present from birth and is, usually, hereditary. So a child showing poor ability, as measured by the tests of mental ability, cannot be trained out of his stupidity. Apparently each child inherits a certain capacity to learn, and his proportional ability remains roughly the same throughout life; if a 6-year-old child is the dullest out of one hundred children of his own age he will still be, roughly, the dullest of the hundred when 16 or 60. Finally, it appears that mental ability develops or "grows" in much the same fashion as does the body as a whole; and as physical growth ceases along in the 'teens, so does the growth of mental ability stop. We may suppose that mental maturity is reached by the average individual somewhere between 13 and 18.

These last statements may seem to the reader somewhat difficult to believe; still the mass of evidence to date supports them. The usefulness of tests of mental ability in the schools, however, is by no means bound up with these theories. If a teacher will only think of these tests as measuring, in general, ability to do school work, and if she will think of this ability as very little affected by training, she will have the essence of the

notion on which the use of such tests is based. In the present chapter three types of test of mental ability will be discussed. First to be taken up will be various scales for measuring general ability which are given to children one by one — that is, which are individual examinations. Following this, group tests of general ability will be discussed. Finally, a word will be said about tests of special ability. Little has been done along this line, however; most tests of ability aim to measure general mental ability, often called "general intelligence."

2. SCALES FOR INDIVIDUAL EXAMINATION

A number of scales and tests for the measurement of general mental ability by means of individual examinations are now available. There are, for example, at least five revisions of the original Binet-Simon Scale, all by different psychologists. However, the one scale which is so much more commonly used than all the others together as to be *the* type of mental ability tests for the psychological layman, is the Stanford Revision and Extension of the Binet-Simon Scale. So the present brief discussion will be concerned, in the first place, entirely with this scale.

General nature of the scale. The general nature of the tests included in the Stanford-Binet Scale has already been indicated. The examination consists simply in asking the child very "common-sense" questions regarding the difference between a stone and an egg, the meaning of "pity" and "justice," the proper course of action if one is going to school and is afraid he may be late; the child also is required to repeat series of num-

bers after the examiner, to give from memory the sense of a newspaper item which he has just read, and so on. The tests are arranged in order as they are passed by children of different ages; thus the memory passage just mentioned is satisfactorily remembered by about three quarters of 10-year-old children, and so it is considered a "10-year" test. The exact nature of the scoring need not be gone into here. It need only be said that if a boy passes as many tests as the average 10-year child, he is considered to have a "mental age" of 10 years; if he passes only as many tests as the average 7-year child, the boy is said to have a "mental age" of 7 years, and so on. The scale thus expresses degrees of mental ability in terms of age of the normal child possessing those degrees of mental growth.

Evidently a 10-year boy with a mental development of only 7 is somewhat subnormal; and a 10-year boy with a mental age of 13 is unusually bright. The Stanford-Binet Scale provides for a further expression of the ability of a child in terms of per cent of that which is normal for his age. So the 10-year boy with a mental age of 7 may be said to have a 70 per cent mentality; he shows a mentality 70 per cent of what he should have. This percentage statement is called an Intelligence Quotient, or IQ. Similarly the 10-year-old with a mental age of 13 has an IQ of 130.

Practicability. The directions for giving and scoring the Stanford Revision are presented by Professor L. M. Terman (the deviser of the Stanford Revision) in a book of some 350 pages, entitled "The Measurement of Intelligence"; a brief manual of directions may be had, but is hardly suitable for an inexperienced examiner. Besides this book, the Binet equipment consists of a

set of cards on which appear various pictures, figures, and other materials used in the examination, and a record booklet. A set of weights and a "Healy Puzzle A" are also desirable but not essential.

As may be inferred from the size of the manual of directions, it is not easy to learn to give the Stanford-Binet Scale; much study and a considerable amount of practice are necessary before an examiner may be considered sufficiently skillful to make her results reliable. As has been said, only one child can be examined at a time; the average examination requires between 30 and 45 minutes. Examining should be done in a quiet room, in which there is no one besides examiner and child. As may also be inferred from the size of the manual, scoring is intricate, involves many special rules, and is intimately tied up with special directions for giving.

The Stanford-Binet Scale is thus somewhat expensive so far as the first equipment is concerned; and it requires the expenditure of much time and effort in learning to use it properly, and in the examining. It is thus much less practicable than the group scales for measuring general ability now available. However, for certain purposes it is of great value.

Use. In order to understand the distinctive contribution of the Binet Scale to the solution of certain educational problems, the nature of the scale from the point of view of the child should be understood. For the child, the examination is delightfully informal, and seems more like a pleasantly diversified conversation than a real test. Further, it is the business of the examiner to keep the child interested. So it is possible

with the Binet Scale to measure the ability of children who are emotionally unstable, and incapable of applying themselves steadily to an uninteresting task, much more accurately than by any other method. It is also possible for the examiner, in the course of such individual examinations, to observe the child's general character very closely. The scale is thus especially valuable in the study of "problem" children — children who are temperamentally peculiar. The examination also involves, in the early ages, no reading or writing, and is sufficiently interesting to hold the attention of young children. So it can be used with children in the kindergarten and first grade, where group tests are comparatively unsatisfactory. Finally, it may be used in intensive study of very dull and very bright children, to supplement and verify results with group tests of general ability. The Binet Scale should be used, then, (1) in the study of temperamentally peculiar children, (2) in the study of very young children, and (3) to verify the results of group examining on very dull and very bright children. The applications which may be made of such results to school practice will be referred to in the following chapter.

The Herring Revision of the Binet-Simon Tests which has only recently appeared is very similar to the other revisions in its purpose and the type of questions asked. It embodies, however, some innovations in technique. For example, a preliminary estimate of mental age is made on the basis of the first four tests. This serves as a guide to the economical administration of the remaining tests. The common all-or-none method of scoring is replaced by a method which gives partial credit for partial answers. By virtue of this and other improvements of technique, the

administration of the test is much shortened and its reliability increased.

Other tests for individual examinations. A word remains to be said with regard to other tests for individual examinations. Of these the most important are the performance tests. A variety of such tests have been devised. They consist for the most part of simple puzzles. A very simple test of this type consists of the figure of a man cut out of a thin piece of wood; the legs, arms, and head are cut out of separate pieces, and the task of the child is to put these pieces together to make the figure. More complicated puzzles require the putting into a frame of blocks which will just fill the space, and that only if put together in exactly the correct way. Still others of these tests consist of pictures with parts cut out; and it is the problem of the child to put in each space the piece having on it the picture of the object which best makes sense with the picture as a whole. The best of such series of performance tests is perhaps the Pintner-Patterson Performance Scale.

Such tests are of great value in testing children with speech defect, children who do not speak English readily, or other cases who cannot do their best on a scale which is largely verbal, as is the Binet. The tests are easily given, and fairly easily scored. Unfortunately, however, the puzzles and other materials are rather expensive. So the use of such tests is confined largely to clinics, where such special equipment can be afforded.

3. SCALES FOR GROUP EXAMINATION

As has been said, to learn to give and score the Binet Scale satisfactorily requires a very considerable amount of study and practice. What is worse — from the point of view of testing economy — the scale can be given to only one child at a time. The performance tests, besides being expensive, are also very uneconomical as regards time. Under such circumstances, the devising of group tests of mental ability was naturally suggested.

The construction of satisfactory group tests involved, then, provision for the answers to be indicated on paper instead of orally as is possible in an individual examination, provision for the simple and rapid recording of answers by those examined so that the time of writing would be negligible in comparison with the time for thought about the test on the part of the one tested, and provision for the rapid scoring of answers by unpracticed scorers in such a way that all scorers would give exactly the same credit for the same performance.

These requirements were met by Otis in his first group test by the expedient of providing on the test paper alternative answers to each question, only one of which was correct. This device was adopted by the army for use in the army group tests, and having proved successful, has been adopted almost universally in the large number of group scales which have been recently compiled for use in schools.

General nature of group scales of general ability. Evidently such tests must put before the children, on a printed blank, problems of the same general nature as those appearing in the Binet Scale. That is, the tests must not include matter specifically taught in school; they must, nevertheless, require the exercise

of "intelligence" — they must involve complex mental processes and ability as great as is required in school work. So these group scales are made up of such matter as the following (quoting from the National Intelligence Tests):

The number of days in a week is 5 6 7 12
The kitten is the young of the dog cat lion sheep
The day before Thursday is Wednesday Tuesday Friday Monday
Cheese comes from butter plants eggs milk

The children are told to "draw a line under the one word that makes the sentence true." Such questions clearly involve general information and avoid specific elements of schooling; they require general ability. The "common-sense" nature of the problems appears more clearly still in such a test as the following:

Do flowers bloom? ..... yes no
Are apples good to eat? ..... yes no
Are some houses built of stone? ..... yes no
Is the sky ever gray? ..... yes no

The children are told to "read each question and draw a line under the right answer." Another test presents such items as the following:

elephant (circus ears hay keeper trunk)
mouse (back cat eyes cheese trap)
hoe (blade digging garden handle rust)
iron (coldness polish rust strength weight)

The directions are, "Draw a line under each of the two words that tell what the thing always has."

The general and common-sense nature of these questions is obvious. They require general ability, practical judgment, mental alertness; they do not require schooling in history or geography, or other

special training. The modern group scale of general ability consists of a booklet containing anywhere from four to ten such tests. The great problem in building these group scales is to involve a sufficient amount of material to give a reliable measure; and it is particularly in the effort to put into these examinations a large number of questions without increasing unduly the time required to take and score the scales that the special features of the present-day group test, which make such tests seem at first glance extremely artificial and even fantastic, have been developed. That is, in these tests directions are cut down to the very briefest minimum; answers are indicated by underlinings or checks instead of writing, and scoring is done by stencils permitting very rapid count of the number of marks correctly placed. As a result of these special features the present-day group scales are, in spite of the fact that they may include 200 questions or over, thoroughly practicable instruments for use in the public schools.

Practicability. Giving these group tests is almost always extremely simple; all the best group tests can be given by the average classroom teacher after a very brief study of the manual of directions. In fact, certain scales now available for use in high school require little more of the examiner than that he should pass out the blanks and then collect them again at the end of 20 or 30 minutes. The taking is easy for the children; the newest tests do not require of them more than 30 minutes of interestingly varied work. Finally, scoring is straightforward and objective; the best tests require nothing more than that the teacher should count up

the number of words correctly underlined or crossed out. These scales, then, are most striking examples of the test builder's "art." As a result of inventions in this field, it is now possible to give a 200-question examination after a 10-minute preparation; the children answer these 200 questions in 25 minutes; and these 200 questions may be scored in not more than 3 minutes. Not all of the group scales are thus far developed, but the usableness of the best group scales is remarkable.

Use. The results yielded by the group tests have very much the same significance as have the results with the Binet Scale; a high score means a good general ability and a low score a poor general ability for school work or anything else the individual may undertake. However, because of the large numbers which may be tested by means of the group scales, more services may be rendered by them than could be rendered by the more time-consuming and less practicable individual examination. Briefly, it may be said that these group tests may be employed for three purposes: (1) They may be used to study the accuracy of grade and section placement, and to improve such placement. (2) They may be used in educational guidance, and (3) they may be used to yield a measure of "pupil material." These uses of the group test will be considered in greater detail in the next chapter and need not be gone into here.

It must not be supposed that group tests of the type above described are capable of as accurate and unequivocal a measurement of mental ability as the individual test. In the first place, the tests just described involve the ability to read; and they are for the most part somewhat

artificial in nature, and so may bewilder a child and lead him to make mistakes, not because he did not know the answers to the questions, but because he did not understand the directions sufficiently to know just what he was to do. In the newer tests, however, these difficulties are of comparatively slight importance. The various problems are very briefly and succinctly presented; there is, thus, only a small amount of reading in proportion to the amount of work called for; and the test form hampers a child's performance very little in these newer tests, partly because the directions are very clear and the questions presented in a very direct manner, and partly because it is becoming common to have practice exercises which habituate the child to the form of the test before he begins to work on that part of the test which counts in the score.

The objections to group tests of mental ability based on the general contention that they are invalid because they are too much conditioned by ability in silent reading, or are so artificial and involved as to be primarily tests in following directions, are thus relatively unimportant. The more recent tests are little subject to this type of criticism. There is, however, one disadvantage to the group tests which should not be lost sight of, and which applies to group tests in the school subjects as well; it is an objection especially serious, however, when measurements of general ability are attempted. As was mentioned (when discussing the Binet Scale) it is the examiner's business, in giving the Binet Scale, to keep the child interested. With the Binet, then, it is possible to obtain a true measure of ability even with those children who, because of emotional instability and lack of steady attention, do not apply themselves well to school work but who nevertheless are fairly bright. Such children sometimes do poorly on a group examination of general ability because of failure to apply themselves attentively during an examination. This point should be kept in mind in interpreting results obtained by group tests of mental ability. It is a

point of less importance than one might suppose, however, because such tests are distinctly interesting and so hold the attention of many of the unstable children.

When there is suspicion that certain children have scored lower than they should, because of such factors as those above mentioned, these children may be given individual tests. It should be mentioned, however, that even in checking on these special cases a group test may be employed. If these children did poorly because of nervousness, or because of the unusual nature of the examination, they will do distinctly better on a second trial. Or, the examiner may go over the group examination of such a child with that child, asking some of the questions orally and trying to find out just what influences may have caused the child to do poorly. Usually such special individual investigation will show that the child did poorly because he was dull — that is, the first rating will be vindicated!

A surprising number of these group examinations for measuring general mental ability have been issued during the past three or four years, and new examinations are constantly appearing. Under the circumstances, the schoolmen interested in such work are often at a loss as to which examination they should use. A statement with regard to the comparative merit of the examinations must, then, be based on careful study of the test materials and the available published matter with regard to them. Special thought should be given to the matter of practicability; the most usable examinations are the ones to use.

Among these various group scales are the National Intelligence Tests, devised by a committee composed of Doctors Haggerty, Terman, Thorndike, Whipple, and Yerkes. The tests have been developed, thus, by a group of men who are unquestionably leaders in this field. These

tests are based upon experience with the army tests, and they involve new and very valuable features besides. There are two parts to the examination, each part consisting of five tests. It is thus possible to use either Scale A or Scale B as a brief examination for rapid survey work. However, for a more reliable rating on individual cases it is advised that Scale A be given one day and Scale B on the day following; any special circumstance affecting results on one day will thus not affect the entire examination. For each test there is a practice exercise which familiarizes the children with the test before they begin work on that part of the material which is to be scored; a slowness in habituating to the special test form thus plays little part in the results. The general get-up of the examination is admirable, the directions are clear and full. Duplicate examinations are being issued from time to time so that these tests may be used over again and again with the same children without employing the same material; opportunities for coaching, or gains from previous experience with the tests, are thus largely obviated. The National Intelligence Tests are intended for Grades 3–8.

Several other examinations are now available for use as measures of general ability. Among these are the Terman Group Tests of Mental Ability, for Grades 7–12; the Miller Mental Ability Tests, for Grades 7–12 and college students; the Otis Self-Administering Tests of Mental Ability: Intermediate and Higher Examinations, the first for Grades 5–8, and the second for Grades 9–12 and college students; the Haggerty Intelligence Examination: Delta 2, for Grades 3–9; and the writers' "Cross-Out" Scale for Grades 3–12.

Tests for children who have not yet learned to read. So far the discussion has dealt with tests for use in the upper grades. Group scales of intelligence which can be given to children in the first three grades — the children who have not vet learned to read sufficientlv

to make possible the employment of the examinations above described — are of still more recent appearance. The general nature of these tests can be briefly described. In general they consist of pictures, geometrical forms, or dots, which are printed on the blank instead of words. Thus, in one test there are a number of pictures of objects and in each picture one part is drawn wrong (one item of this test shows a letter with the stamp in the wrong corner; another shows the flag with the field of stars in the lower instead of the upper corner). In each case the children cross out the part that is wrong. In another test the children are shown pictures, each of which presents several objects; they are to cross out the one object that is different from all the others. Tests of this sort require as complex a type of ability as is required for the mastering of first-grade work, but do not involve any special ability — such as reading or writing — that is developed in the school. All that they require of the child is that he should be able to hold a pencil so that it will make a mark. These tests are all fairly easy to give, though young children are somewhat hard to work with because of their restlessness. The children find the tests interesting and easy to take; and scoring is extremely easy. The tests are, then, practical instruments for use in the schools, and it may be expected that group tests for the younger children will soon be in general use.

Among the group scales for young children are the Detroit First-Grade Intelligence Test; the Otis Group Intelligence Scale: Primary Examination; the Haggerty Intelligence Examination: Delta 1; the Kingsbury Primary Examination; and the writers' Primer Scale.

Such work with very young children is still too new to make possible any very definite conclusions with regard to the comparative merits of these different scales. Nor is the significance of the results obtained altogether clear. The children are so young that testing them at all by the group method is something of a task. They do not work well together, they do not apply themselves well, they make themselves work in the mere making of a simple mark, they copy the work of others if not prevented from so doing, and, all in all, require a very different procedure in examining from the sort of procedure that will work very well with older children. In fact, the primary children demand in testing, just as they do in teaching, special methods. Especially is it necessary that tests in the primary grades should progress because of interest, not because of school discipline. The examiner must be sure to make his directions not only clear, but of interest so that he will catch the attention of the children and cause them to apply themselves; any attempt to proceed in such a manner as to make the disciplining of the class an obtrusive factor is not likely to be successful. However, if the examiner is reasonably careful and tactful, such tests may be considered to yield results of distinct value in the comparison of groups and of a considerable reliability in the measurement of individual children.

4. TESTS OF SPECIAL ABILITIES

So far in the chapter tests in general ability have been presented and briefly discussed. Tests of special ability remain for brief mention. Such tests have as

yet been little developed. It seems probable, however, that materials of distinct value for educational and vocational guidance of the older children may soon be available. In the present chapter only brief mention can be made of efforts in this direction.

Of chief interest to school people are tests aiming at the "predetermination" of ability or disability in a certain type of school subject. Thus, a recent scale attempts to determine the probable ability of a child to learn a foreign language. The tests are the result of an analysis of the abilities which appear to be involved in the learning of a foreign language. The scale is intended to be given to pupils who have had no work in foreign language, as a means for determining whether or not they may successfully learn a foreign language. If such a test can save pupils the waste of time and the discouragement incidental to failure in language work, and can direct them instead into studies more suited to their abilities and interests, its use is surely more than justified.

Another set of tests endeavors to determine ability in mathematics. Still another series investigates mechanical ability; and efforts are being made to investigate ability in business, in salesmanship, in clerical work, and so on. Strictly vocational tests, however, hardly come within the scope of this book. It should be clear that vocational tests are of distinct interest to the high school principal, as many of the children in the high school are in need of vocational guidance; and educational guidance is becoming more and more important in high school work. Anything

that tests can do to help the counselor in this type of work should be carefully considered. Work along these lines is only just beginning, but great advances may be expected within the next few years.

CHAPTER TWELVE

Use of Tests of General Mental Ability

1. LIMITATIONS OF SUCH TESTS

IT will be well, perhaps, before entering into a discussion of ways in which tests of general ability may be used in a school, to attempt further definition of the nature of such tests and their limitations in dealing with educational problems. We have spoken vaguely of general ability somewhat as though we might mean simply ability in general, and the question as to just how much a school should be interested in the "ability in general" shown by its children has been only touched upon. Further, nothing has been said as to whether ability to learn arithmetic may not be different from ability in language or manual arts; and the question, chronically prominent in classroom experience, as to whether a child's interest and *will* to learn may not be quite as important as that child's mental ability has been avoided. Before discussing ways in which tests of mental ability may be used, it is surely necessary to take a stand on these questions.

What is "general ability"? First comes the question as to what, from the point of view of the school, general ability may be. The writers believe the teacher will find it most profitable to think of general ability as meaning, so far as she is concerned, *general ability to learn* — general ability to profit by instruction; and (to anticipate a point which will be returned to shortly) the great problem in adjusting school organization to

the individual child is to give each child opportunity to learn in proportion to his ability to do so.

General ability may very likely involve other elements than ability to learn in school. However, for the teacher this is the important element; and she will be saved much loose thinking, and lose very little from the concept, if in her study of the matter she will consider "general ability" as meaning "general ability to learn." The question then at once arises as to whether learning ability really is general. That is, are the children who do well in arithmetic also the children who do well in spelling, language, drawing, and the other subjects in the curriculum?

Is ability general? The problem can be touched on only briefly within the limitations of the present chapter. It should, perhaps, be mentioned first of all that grade organization in the elementary school implies such general ability; and, in so far as promotions are made by grade instead of by subject, school organization seems to be subscribing to the doctrine of a general capacity. However, promotion by subject is being widely urged, and departmentalization is extending farther and farther down into elementary school work. Further, in the chapters on tests in the school subjects, there has been constant reiteration of the statement that learning is very specific; it has been said, for instance, that multiplying six times nine is not quite the same thing as multiplying nine times six. And, to return to a question mentioned in the first paragraph, every teacher knows children who draw and write beautifully but are failures in arithmetic; she also knows boys who do well in arithmetic but are poor in their language work.

Those children who do well in writing and drawing, but poorly in arithmetic and reading, offer no difficulty to advocates of the theory of general ability. It is well known that the feeble-minded may do beautiful embroidery, may show good technique (though no interpretation) in playing musical instruments, and may write and draw with great skill. Complicated manual tasks may be learned by dull children; they learn them very slowly, but they can learn them. The feeble-minded child shows himself inferior especially in the more complex subjects requiring some original thinking, and in those subjects requiring the use of ideas and symbols — that is, in such subjects as reading and arithmetic. But how about those boys who do well in arithmetic but poorly in their language work?

It may be said shortly that such cases are more rare than is sometimes supposed; and when they do appear they are very likely to be the result of special interest rather than of special ability — perhaps the result of early home training, or of enthusiasm along a particular line shown by a parent or teacher. Interest along a different line might have produced a different "ability." It should also be noticed, in this connection, that the presence of unusual attainment in a given field may be due to accidents of instruction; therefore, that (to return to an example just mentioned) a child is able to multiply six times nine more quickly with less frequent errors than nine times six by no means implies that he has a special ability for the six-times-nine combination. This combination has simply been drilled on more, and more frequently met.

It may, then, be concluded that ability is quite largely general, and that, *other things being equal*, the child who does well in arithmetic will do well in reading, history, and geography also. (He may be poor in writing and music, but that is a different matter.) Our tests of general ability may, then, be used to indicate general capacity for school work; and it can readily be realized that a measure of a child's total "capacity for profiting by instruction" should be of vital importance to a teacher in adapting her instruction to fit individual needs, and in guiding a child, educationally, into the type of work for which he is best fitted.

It should not be inferred from the above discussion that there are no special abilities. Occasionally children appear who are gifted along a special line. It should also be realized that, though innate ability is for the most part general, special ability tests may be, nevertheless, desirable. Thus we may wish a test to discover whether a child has ability for the study of modern languages. This should not be taken to imply that specific ability for learning these languages is born in a child. Nevertheless, by the time a child is ready to enter high school he may have developed a considerable aptitude in this particular direction, as a result of good training in English grammar, wide reading, and an interest in language work, together with a special interest in and knowledge of the history of the particular country whose language he wishes to study.

The importance of interest and application. One important qualification with regard to the dominance of general ability remains to be mentioned, however. Two children, equal in general ability, may nevertheless differ decidedly in the extent to which they do actually profit by instruction, as the result of differences

in interest, application, and willingness to work. One might almost say that the two great factors in school achievement are ability and application. In general, the brighter children are the children with the more lively interest, and are the best behaved; but if the school fails to provide educational opportunity in proportion to the extra ability of these brighter children, a most unfortunate disciplinary problem often arises. The child is kept in a grade where the work is too easy for him; so he has little to do, develops indolent and mischievous habits, and may be even expelled as incorrigible.

So much then, briefly, with regard to the nature and importance of general ability. The services which may be rendered by tests of general ability must now be pointed out.

2. USE OF INDIVIDUAL TESTS IN THE SCHOOLS

Individual study of unusually bright and unusually dull children. Use of tests which are given to the children individually requires a great deal of time and considerable skill on the part of the examiner. Evidently such tests should not be used except on special occasions, in dealing with special cases, or where group tests are inadequate or not applicable. The most common use of the Binet Scale is to investigate backward children, to determine whether they may be feeble-minded. Perhaps the advantages of such an investigation are sometimes overestimated. It is hardly worth while to find out that a child is feeble-minded unless something can be done about it. It should be understood by the teacher that state institutions for the

feeble-minded are always crowded, and cases which are a social menace because of delinquency are always given preference in admission. The schools can find little relief from the burden of the feeble-minded by disposing of them in this way; only when a child is also delinquent is there any hope of such a solution in dealing with a problem case — and then it may be months before there is any chance of admission. It may assist a teacher somewhat to know that a child is a mental defective; but unless the school system has classes for backward children, or some other means for dealing with these cases, a Binet examination to determine the degree of mental defect will be of little service. Teachers all too frequently seem to feel that a Binet test will in itself assist in the situation. It cannot be repeated too frequently that tests should not be given merely for the sake of testing; the test should be given as an aid in the improvement of instruction, and if no means are available for the necessary improvement in the system, use of the test is hardly worth while. It is generally recognized, at present, that continued failing of the backward children, and their retention in the first three or four grades, is unwise. Unless it is possible to put these children into special classes or to arrange a special course stressing manual work, careful diagnosis of these defectives by means of time-consuming Binet examination is hardly worth while. The time and effort involved had better be put elsewhere.

Much more worth while is special study of the unusually bright children, for special treatment of these bright children is much more easily obtained in the

average school system. Instead of trying to find the feeble-minded in a school, a principal should try to find the gifted children; and effort should then be made to accelerate these children in their school work. There are few schools, at present, which do not contain a number of children two grades below the grade in which their ability would properly place them. Six weeks spent in summer school, half a year in a special "rapid progress" section, or a little extra outside help from the teacher would make it easy for these children to skip half a grade — or more. The importance to the child of saving a year in his elementary school work can hardly be overestimated; it means a year added to his later life. Incidentally, such acceleration means a saving to the school; it means a year less of expense in educating that child. If the Binet Scale is to be used in finding those children who are at the extremes in ability, the *upper* extreme should by all means be selected for study.

Individual study of the temperamentally peculiar. It has been urged that a study of dull children by means of the Binet Scale is often unprofitable. One exception should be made to this statement. The child who is backward in school and also emotionally unstable, or otherwise peculiar in temperament, may need study by means of the Binet examination because of the special opportunity which this scale gives for the control of such emotional factors during the examination. Emotional instability is very often found with mental defect; the mentality is too weak to control the impulses. Such a child may be vicious; and if he is both vicious *and* feeble-minded, it may be possible to secure his

admission to an institution for the feeble-minded. It may be, on the other hand, that the unstable child is fairly bright but unstable because of some special physical condition, such as chorea. Medical attention is then advisable. Or it may be that the emotional difficulty is environmental in origin; there is some difficulty at home, some incompatibility with the teacher, or other special factor — and the child is, nevertheless, capable of doing good school work. Solution of the emotional problem is then to be sought. It may be, finally, that the school is responsible for the lack of interest in school work shown by the child. Perhaps the child is unusually bright and is in a grade where the work is not sufficiently difficult to keep him interested. Under such circumstances the Binet examination may contribute something of value by pointing out the superior ability of the child in spite of his inferior school work. An extra promotion, or even some extra work, will sometimes accomplish wonders. The writers know of a case of persistent truancy which was entirely cured by a double promotion made on the basis of a Binet examination.

Individual study of young children and of special cases. The Binet Scale may, then, be used in dealing with the subnormal, the brilliant, and the unstable. It remains to be mentioned that there are certain special situations in which such an examination as the Binet, or a series of performance tests, is needed because of illiteracy or language handicap. In certain school systems it is now the practice to give all the first-grade children a Binet examination soon after entrance to school; the children are then placed in "fast" and "slow"

sections according to ability. The Binet Scale is used with these young children because of the difficulties, already mentioned, in testing these children in a group. With the development of better group tests for young children, such use of the Binet Scale will become largely unnecessary. In schools where there is a large foreign element, use of the Binet Scale and of performance tests is sometimes desirable. In study of children showing some special defect, — such as speech difficulty, — intensive study of the child, using individual tests, may be needed.

To summarize, then: Individual tests of ability may be used (1) to study children at the extremes of ability — the unusually dull or the unusually bright, (2) to analyze the difficulties of the emotionally unstable children, and (3) to deal with cases to whom group tests cannot well be given — as very young children, children unfamiliar with English, or children suffering from some special handicap.

3. USE OF GROUP TESTS IN THE SCHOOL

Measurement of the accuracy of grade and section placement. It may be said at once that three uses may be distinguished for group tests also. In the first place (1) the tests may be used to study the accuracy of grade and section placement in a school. Such use gives a remarkable insight into the educational problem with which a school is wrestling, and may be made the basis for a highly profitable readjustment. For such a study a group mental ability examination should be given throughout the school and the results tabulated keeping the grade, sections, or other divisions

separate. Study of the resulting table is often most illuminating. A few children testing remarkably above the grades in which they are placed will probably be found; a child in the fourth testing above the median for the seventh, for instance, may appear. A special effort to double-promote this child and get him into a grade where the work is more proportional to his ability is then in order. Before doing this a Binet examination may be desirable — though in grades above the fourth the group test may usually be considered sufficiently reliable for such procedure.

Such extreme cases should be studied first. But there should be the same consideration of cases less strikingly misplaced. A general truing up of the grade divisions should be attempted; so far as possible each child should be in the grade in which he is best able to do the work. Division into "fast" and "slow" sections should also be made more accurate; and if there are special or ungraded classes, the personnel of these classes should be studied. If such a survey is made in a number of schools, the superintendent may find it profitable to compare the accuracy of placement in the different schools; he may, for instance, count up the number of children in each school who test two grades above or below the grade in which they are placed. Such a tally should give a rough measure of the extent to which a school is, or is not, taking account of the differing degrees of ability among its children.

It must not be understood that the writers advocate any arbitrary moving about of children. There should always be careful study of the total situation; and always there should be thoughtful consideration of the teacher's judg-

ment in the matter. As has been said before, the tests should not take the place of the teacher's judgment, but should serve to aid that judgment. It should also be remembered that (as stressed in the chapter on statistics) test results are somewhat inexact and rough. If a child in the 5B grade tests above the median for the 5A grade, when the medians are only a few points apart, the fact may be of little significance. Only when a child tests considerably above or below the grade he is in — one, two, or three grades, or more, away — do the findings acquire great significance. Similarly, sections should not be made up altogether on the basis of the tests. Children testing in the lower quarter of a class probably belong in the "slow" section, and those testing in the upper quarter in the "fast" section; but children testing close to the median should be assigned to one section or the other largely according to the judgment of the teacher, after she has considered both her past acquaintance with the children *and* their test score. Always, any judgment with regard to the health of a child, as conditioning his ability to stand the extra work required for a double promotion, and any knowledge on the part of the teacher concerning a pupil's preparation to date, should be taken into account. Always — to repeat — tests should be used as an aid to the teacher; they should never be used to overrule or antagonize her.

Group tests as an aid in educational guidance. Tests of mental ability may, then, be used to great advantage in improving grade and section placement. Closely related to this is a function most prominent at present in senior high school, but also desirable in junior high school and in the grades as far as possible. Group tests of intelligence may be used (2) in educational guidance. And educational guidance, it should be realized, is really one phase of vocational guidance; for the child's schooling is his work, and it should fit him for the work

he will do in later life. Obviously, his choice of a life work should be dependent to a certain extent upon his ability; and the school should train him along the lines for which he is best fitted. Tests of special ability would seem desirable, but such tests are as yet little developed. Meanwhile much may be done by taking account of a child's general ability. The less intelligent children should not be advised to take college preparatory work; the gifted children should be especially urged to obtain such an advanced education. The dull children should rather be urged to elect courses of a fairly immediate vocational value — courses in commercial subjects or in the trades. In the junior high school these children should be given work in practical manual training or in the domestic arts. It has been found that such guidance tends to cut down the number of failures by keeping the duller children away from such subjects as Latin and algebra, in which they would be likely to fail; it holds the children in school, and off the streets, longer because they are given work they can do, and keeps such children contented and happy in their school work so that they remain in the school atmosphere during the important period of adolescence.

Measurement of "pupil material." For the superintendent or supervisor tests of mental ability may yield a measure that is of special interest. Such tests may (3) give a measure of the "pupil material." A superintendent may well find that the children in one of his schools average more than a year in "mental age" below the children in another school. If the children in the first school are thus dull, it is clear that the

teachers of this school have a difficult problem to meet, and it should not be expected that results on tests in the school subjects would be as high as in the second school. In fact, a superintendent simply cannot rightly understand the problems faced by his various schools without such findings regarding the "capacity for school work" which the children in these schools possess. It should also be mentioned that differences in average are not the only differences which may appear. There may be differences in the range of ability also; one school may contain *both* very bright and very dull children while another school may have only inferior or average "pupil material." It is evident that special classes or ungraded rooms are more needed in the first school than in the last. It deserves to be mentioned, finally, that such a complete survey once made need not, for this purpose, be repeated every year. Such differences from school to school are due to differences in the type of people living in the different neighborhoods and are relatively unchanging over a period of years, unless some radical change takes place in the population of a district.

It remains only to be said that the above list of uses for tests of mental ability is by no means complete. It will serve, however, to make clear general methods for employing such tests in dealing with school problems. The discussion has been made relatively full because use of these tests in the schools is a comparatively new thing, and mistakes are frequently made. It is believed that the above suggestions, based on extended experience with such tests, will prove of value.

PART FOUR

IMPORTANT GENERAL PRINCIPLES REGARDING TESTS

CHAPTER THIRTEEN

How Tests Are Made

IN the first section of this book those elementary facts were presented with regard to the nature of tests and ways in which they may be used, that were felt to be essential for those who are just beginning to use them. In the two succeeding sections tests in the school subjects, and tests of general ability, have been described. This final section aims to outline certain further matters with regard to the making of tests, and the formulation and carrying out of a coherent testing program. The present chapter will attempt to make clear the various steps through which it is necessary to go in building a test; some understanding of the processes involved in test building would seem to be worth while as a background of information for the user of tests. In the progress of constructing a test, four distinct steps may be seen: (1) The originator must first carefully define his problem. (2) He must then decide upon the form that he will consider satisfactory. (3) Then begins the weary task of making up the items for first trial and the selection of those items to be included in the final form. (4) Last of all comes the validation of the test and its extended use in final form so that adequate norms will be available. If the test under consideration is a single test, or scale, these four steps may be considered inclusive of the fundamental processes involved in test construction. However, if the originator is engaged in building an examination which includes several separate tests, there still remain (5) certain special problems involved in the

combining of these different tests to make the most satisfactory total examination. All the steps outlined above are necessary and will be taken up in some detail by means of an example which will be used for greater clearness.

1. DEFINITION OF THE PROBLEM

Necessity for a clear formulation. It is evident that all further steps are conditioned upon a clear formulation, in the first place, of the problem with which the test is to deal. One must know with some exactness just what one is trying to measure; it is not enough that one set out with a vague notion of measuring, for instance, "spelling ability" in general. The exact object of the test must be clearly apprehended from the start if the final test, which is the outcome of the work, is to yield results sufficiently definite to result in the "improvement of instruction" that should be the ultimate aim of testing. In general, the more specifically and definitely the test problem is formulated from the very first, the more directly useful will the test be to the users. Thus, a test builder who sets out to test "ability in English" without any more specific limitations upon the field of his research is not as likely to make a test useful to teachers as that made by the man who builds a test to measure the ability of children of a certain grade to use quotation marks; if the test is to be a "general" test, all the elements to be included should be carefully taken account of. It should be clear that an understanding of the object of the test — in a detailed rather than a vague, general way — is the first step toward a good test.

Determination of certain minor aspects of the "problem." It is also necessary to decide in advance of any actual work the range of applicability of a test; if it is to be given to high school pupils, the items, from the very start, will be formulated differently from the items of a test intended for younger children. If the test is designed for use in the primary grades, several difficulties beset the test builder that do not function in the upper grades. Further, one must settle in advance whether or not the test is to be given and scored by teachers. If it is, and most tests are so intended, the originator must make plans for a cutting down of any clerical work, for the adoption of directions suited for use by persons relatively unskilled in testing methods, and for a recording of the answers by the children in such a way that the scoring will not consume an inordinate amount of time. If such details are not foreseen and worked out with care, the resulting test is likely to be less suited for ready use by teachers than it might otherwise have been. One should have also clearly in mind the probable statistical handling of the results with reference to the particular practical questions which the test is to aid in solving, so that the arrangement may be as convenient as possible for such handling. Many tests require unnecessary work of those who use them simply because the originator did not foresee the particular statistical methods which those using the test would wish to employ. This entire process of formulating the problem of a test, and of settling these various matters of detail, may best be illustrated by considering an actual problem. It must not be understood, however, that the particular test

presented is absolutely typical of all tests, for it is not; but it illustrates very well the first formulation of a test with reference to the problem. Reference will be made from time to time to other types of test construction.

The speed test in silent reading mentioned in Chapter 8 may serve as an example. In this case the aim was to construct a test which would give an indication of the rapidity with which a child could read very simple matter; from this rapidity one might be able to deduce certain facts with regard to the particular difficulties with which a child was struggling — especially with regard to the persistence of "oral reading habits," which greatly retard the speed of assimilative reading. It was, therefore, desired that the test should not involve any reading matter which was in itself difficult; the score should be conditioned primarily by the rapidity with which a child was able to glance along the lines and assimilate the general gist of what he was reading. Thus any difficulties of vocabulary over which a child might pause should be eliminated; otherwise the children with the more limited reading vocabulary might "hang up" on an item, not because they could not read it rapidly, but because they could not read it at all. The test should, therefore, test the ability of the children to cover a page of reading material rapidly, with a fair assimilation of the meaning; it should specifically point out those children who, because of oral reading habits, are unable to progress rapidly, and should mark these children off definitely from those who have freed themselves from such habits.

Since the transition from oral to silent reading takes place largely in the second, third, and fourth grades, this test should evidently be made applicable to these grades. The first grade appeared as too early for much of this transition to appear, and the grades higher than the fourth rather too late, except in special cases; certainly the im-

portant period of transfer from oral to silent reading is in these three grades, to which it was decided to limit the test. Inasmuch as the children to be tested were very young, it was realized from the start that the form of the test and the type of directions must be of the simplest. It was intended that the test should be given by the teachers; so simple directions and easy scoring and interpretation of scores were aimed at. And, since the test was intended primarily for a separation of those children who were free from oral reading habits from those who were not, a simple statistical handling of the results, thus to sort the children, was to be sought.

From this rather detailed statement of the formulation of the problem in the case of a particular test, the importance of a very concrete definition of the problem may be realized. It should be pointed out that, had the test had some other object in view from the particular one with which it was concerned, this preliminary formulation might have been vastly different. The general process of defining the limits of the problem as shown in this illustration may be considered typical for all tests, although the details would be different from one type of test to another.

2. SELECTION OF THE TEST FORM

It should be obvious without discussion that the form of a test should be as simple as the necessities of the problem permit. There is more than one reason for this. Certainly the user of the test should not be hindered from frequent use of the materials by their complicated nature. If tests are to be used as frequently as they should, it is essential that they should be so constructed as to make frequent use possible; a poorly

constructed, time-consuming test forms one of the best arguments against a wide application of the test method to educational problems. Moreover, a simple test technique is more likely to give valid results than an artificial and complicated form. Whenever, for instance, a test in reading is so complex in its presentation that the taking of it involves the ability to follow complicated and somewhat irrational directions, it is a question whether the test is really measuring what it set out to measure; that is, poor test form interferes materially not only with the usefulness of a test, but also with its validity.

Elimination of writing. Most of the tests now on the market, unless measuring handwriting, do not call for written answers. The elimination of writing sometimes causes a test to have a somewhat artificial appearance, but it does prevent the inclusion in the test score of a factor irrelevant to the object of the test. Thus, a reading test which demands written answers from children will — especially in the lower grades — yield a measure of the ability to read the selection and the questions *plus* the speed with which the child writes and his ability to express himself in written work; and this speed of writing and facility in written expression have surely no definite relation to the understanding of printed material. Writing is also a great inconvenience in the scoring; but the greatest objection to it is that it cuts across and obscures the significant results of the test.

Since the children do not write, it is evident that some other means of indicating score must be used. The originator usually plans for an underlining of the

correct answer, or a checking or circling, or some other simple indication. Such simple manipulations with the pencil serve to indicate the answer, make the scoring easy, and are equally difficult for all children — and so do not affect the score as handwriting does. The originator must select some such means for the children to use; and the more he can rationalize the use of the symbol he selects, the better for the understanding of the test on the part of the children.

The directions for giving the test. If the test form is simple, very simple directions may be used; if it is complex, the directions must be equally involved. The object of the directions to a test is to explain the test to the children in such a comprehensive way that they will understand *exactly* what they are to do; the directions must be clear enough to obviate the necessity for any questions from the children, and they must foresee any elements of confusion. Yet the directions must be short and simple. In the first place, long and complicated directions demand that the teacher make a lengthy preparation before she can give the test properly; such time-consuming preparation once indulged in is likely to prevent further use of that test — or of any other. Further, the directions must be short and clear, or the pupils will not know what they are to do. Long or involved directions result in only a very hazy understanding of the work to be done, and allow the children to become completely bewildered and to attempt to solve the matter in some way all their own. The directions must be just as simple and clear as the form of the test permits them to be.

The scoring of the test. Evidently, if writing has been

eliminated, difficulties of scoring have been considerably lessened; but complications may still remain if the form of the test is not fairly simple. The tests must be so arranged that the scoring is of a routine nature. Weighting of questions, or other arithmetic in the calculation of the score, is undesirable. The most simple and straightforward method is the best. Cumbersome and complicated scoring directions are simply an offspring of poor test form.

The speed-of-reading test under consideration as an example evidently demanded that the test form be such that the only element measured was speed of reading. It immediately follows that writing could not be used, since a measure of the speed of writing would be included in the measure of the speed of reading. But, since it was a silent reading test, some check must be included to make certain that the children really understood what they were reading; it would never do to have the children simply mark the place in a paragraph to which they had read at the end of 3 minutes, for instance, because some children would merely go down the page without attempting to get the meaning. Since writing was definitely eliminated, some other symbol had to be adopted and the test form so selected that some objective symbol could be easily used. After considerable experimentation it was finally concluded that a series of very simple sentences, each containing one irrelevant word, would make material suitable to the problem. Such sentences as, "The sky is tree very blue," would clearly offer no great difficulty of vocabulary that would interfere with the speed of reading; and they would permit of a very ready and sensible use of an objective means of recording the score. The children who understood the gist of the sentence would see at once that the word "tree" was not a part of it, while the children who did not grasp the meaning would not realize this. And the children might record their comprehension by simply crossing out —

or otherwise indicating — this extra word; further, the elimination of something extra or irrelevant would appear to the children to be such a natural thing that they would have no difficulty in keeping the directions in mind. The number of sentences correctly checked would give a ready indication of the number of sentences the children had been able to read in the time allowed. It would seem that such a test form was sufficiently simple to allow of measurement of the ability it was desired to investigate; it is hard to see how such a simple form could interfere materially with the measurement. One further detail with regard to the mark to be called for remained to be settled. If the children were told to cross out the extra word, some children would waste time in completely obliterating it; if they were told to underline it, some children would place the lines so roughly that scoring would be difficult. So it was finally decided that the children should be told to eliminate the extra word by "drawing a line around it"; if, however, they used any other mark it was to be counted as satisfactory, since the test was to be a measure of the speed of reading — not of the ability to follow directions.

A few other details still remained for consideration. The size of the print had to be decided upon. The size and nature of the blank demanded thought; it was decided that a single sheet should be used, since young children are likely to lose their way in a folder of any kind. Since the test was one of speed, it was essential that the children should not see the items on which they were to be scored until they understood what they were to do with them and were all ready to begin the real work of the test. But young children must be given ample time for writing name, age, and grade; and the explanation of examples must not be hurried. So the lines for writing name, etc., and the examples to be discussed with the class, were printed on one side of the page and the sentences of the test proper on the other. Thus, in giving the test all writing and explanation are done at whatever speed the

children can manage conveniently; then, at the command of the teacher, the blanks are turned over and the children start work together on the test itself; when time is up they all turn the papers back while they are being collected.

Such matters of detail in the decisions with regard to test form may seem petty and unimportant; yet they are just the matters that are prominent in deciding whether or not a test is to be widely and frequently used by teachers, and every one of the details above described is important, and must be settled by the test builder. Such details vary of course from test to test, because the nature of the problem so conditions the form; but the essential principles should be thoroughly understood as applying to all tests. This matter of test form has been discussed in detail, because teachers very frequently fail to understand the purpose of the special devices used in tests, and they criticize tests as being artificial. It should be understood clearly, from the above discussion, that such devices have very real reasons for existing; and the purposes of these special devices and their merits should be thoroughly comprehended by all users of tests.

3. CONSTRUCTION OF ITEMS, PRELIMINARY TRIAL, AND SELECTION OF ITEMS FOR FINAL FORM

Selection of material for trial form. As was mentioned in the first chapter, one of the distinctive features of a test is the fact that the materials included are very carefully selected. In many cases such selection begins much before construction of even a trial form of the test; so work on a geography test may be begun by

study of the textbooks in the subject. In other cases available matter is at once made up into an extensive trial form, and systematic selection begins there. But in any instance there must be a trial form of the test.

This trial form contains a great many more items than will be needed in the final test. The test builder cannot tell which items of this trial form will be most satisfactory; but the probability is that at least half of the trial items will have to be thrown away. So this trial form is put together and tried on a large number of school children, with adequate time given for all the children to finish, so that results may be obtained on every item. These results are then very carefully analyzed from every important point of view, and the relative difficulty of each item determined.

Considerations in selecting the final items. The basis for selection of items for a test varies with the problem. The preliminary forms should be tried in a number of localities and a number of school systems and the per cent passing each item studied for each locality in order to be sure that localisms, some feature unduly stressed in some particular textbook, or other special factor does not creep in. The items must be so chosen that the test is not too easy or too hard; there must be (if the test is a "difficulty" test) some easy items and some hard items, so that there will be something suited to all the children to be tested, from the poorest to the best. If the test is a "speed" test, items easy enough for all should be selected and a time limit carefully determined. In so far as possible, each item should be selected according as it contributes to the solution of the problem in hand. Thus, if one is constructing a

test of mental ability, he will select those items which best differentiate very bright and very dull children. Always the items for both the preliminary and the final form must be selected with the greatest care.

In the case of the reading test used as an example, the items were first constructed by selecting sentences from *first*-grade readers, so as to be sure that no difficulties of vocabulary would be encountered. About 30 items were desired for the final test, so 75 were constructed and used in the trial form. No words were used that were not known to be common to first-grade readers. After the first trial items were selected partly as regards difficulty (as will be mentioned shortly) and partly according as the items differentiated those children who were known to have hampering oral reading habits from those who were known to be free from such habits.

Constructing a "scale" in terms of difficulty. The tabulating by item above referred to shows at a glance the relative difficulty of the items. Thus, an item passed by 89 per cent of the children in a given grade is an easy item, while one passed by only 5 per cent is very hard. Having found the difficulty of the various items in terms of per cent, — or, if more accuracy is desired, in terms of Probable Error, — the test builder usually arranges the items in the final form of the test so that they will become increasingly more difficult as the child progresses through the test. Aside from certain statistical advantages of this arrangement which are too technical for discussion here, there is the psychological and practical advantage of allowing the child to begin on the easiest item and then progress as far up the scale as he can. Unless the items of a test are thus arranged in order of difficulty, — the increase in

difficulty from item to item being the same, — it is not properly called a "scale." The term means that the items have been "scaled," or arranged in order of difficulty.

> For the speed-of-reading test, arrangement in order of difficulty was not desired; in fact, tests which measure speed should contain items all of the *same* difficulty. Otherwise, the measure of speed will be confused with a measure of difficulty as well. Sometimes such combined measures are desirable, but for the measurement of mere speed it is essential that the items be of approximately equal difficulty. Therefore, those sentences were selected for inclusion in the final form that were passed by from 75 per cent to 85 per cent of the children in the second grade. When the items of the test were thus selected and arranged, it was found that a test consisting of 32 items with a working time of 3 minutes was most satisfactory.

Such details of selection and arrangement may seem to the lay reader to be unnecessary, but they are really necessary if the test is to be as "efficient" as it should be. The final test will be just as efficient in meeting the problem it is supposed to help in solving as its items are efficient; if the items are poorly selected, the test cannot be satisfactory. The technical details used in the selection and evaluation of difficulty have been purposely omitted in the hope that the general purpose of this step in test construction might be made clear.

4. FINAL TRIAL, STANDARDIZATION, AND VALIDATION

After the test has been constructed from the items of the trial form, after any difficulties in giving, taking, or scoring revealed by the first trial have been remedied, and after the exact timing has been determined upon,

the test is ready to be standardized. This means that it is given to large numbers of children in the proper grades in several cities in order to obtain norms. These results are tabulated and medians — or other measures — are worked out, usually for both age and grade.

This final trial also includes what is usually referred to as the "validation" of the test. That is, the originator started out to test some particular ability; he must now determine whether or not his test, constructed with so much thought and effort, really does measure the ability he wished it to measure. It is one of the disappointing things about this sort of research that what seemed to be a perfectly good test does not always measure what the originator wished it to — and the test has to be "scrapped." The actual procedure in "validating" a test again depends on the type of test and the nature of the problem; but, whatever may be the circumstances, the originator must determine the efficiency of his test in doing the thing he wishes it to do.

The test used as an example was intended to measure the extent to which oral reading habits were persisting and preventing progress in reading among children who should be free from such habits. In order to obtain information on this point, the examiner asked teachers in some of the schools to which the test was given in its final form, to turn in the names of those children who showed the presence of oral reading habits by whispering the words to themselves, following the words with the finger, etc. The scores for the children thus designated were then marked in on the tabulation sheet for their grade in red; and examination of the position of these red marks showed that all these children scored in the lowest 10 per cent for their grade. It was therefore concluded that the test *did*, to some extent at least, differentiate these special cases.

5. SPECIAL PROBLEMS INVOLVED IN COMBINING SEVERAL TESTS INTO A SINGLE EXAMINATION

The previous sections have dealt with the construction of a single test. But some examinations contain several tests. In such cases the test builder must go through the steps outlined for each test, but he is even then not through with the construction of the examination. He must select tests which combine well into a single examination; and he must integrate the different parts of that examination very carefully.

The most common type of multi-test examination is the type used for measuring general mental ability. In selecting tests for inclusion in a group scale of mental ability, the originator tries to use tests which supplement, rather than duplicate, each other. A very common error in such examinations is the selection of several tests of a very similar nature. Instead, each test should involve a somewhat different aspect of ability; each test should make a distinct contribution to the score. So, in making an examination of this sort, tests should be selected that show a relatively high relationship, or correlation, with mental ability, but a low correlation with each other. Such selection of tests requires experienced judgment and involves elaborate statistical technique. The whole matter is too technical for discussion here; but the reader should understand the general problem — the need for a selection of tests which "team" well together, each test playing its own distinctive part in the total result.

The tests must not only be carefully selected, they must be unified into a coherent examination. So far as

possible, directions, scoring, presentations, must be similar from one test to the next. This is important for examinations to be used in the upper grades, but it is essential for examinations for the first two grades. If primary children are told to cross things out in the first test, it is practically impossible to persuade them to do anything else for the remainder of the tests.

Moreover, if one test seems to be more valuable than another it must be "weighted" so that it will count for more in the total score; each test in such an examination as above referred to may be thus "weighted" in proportion to its value. Again the methods used are too complicated for discussion here; but the reader should at least know that there *are* such problems, and should be able to discriminate between different examinations according as they are, or are not, unified into a coherent whole.

Summary. In building a test, then, one must first (1) clearly formulate the problem. One must then (2) carefully work out the test "form." There must, further, be (3) very careful selection of the materials to be included in the test, and careful experimentation with a trial form. Following this there is (4) construction of a final form, and trial of this form, with standardization of the test, and a final demonstration that the test is doing what it was intended to do — is dealing with the problem as first formulated. In addition to all this there must be, (5) when the examination consists of more than one test, a careful unification of the total examination, so that all the tests included make up a coherent total.

It can be appreciated that the building of a test is no

easy task; and at that, the above description leaves out many minor steps altogether. The teacher will not attempt, of course, on the basis of the above brief description, to build tests herself! But it is believed that the value of tests will be much more appreciated if something is understood regarding the way tests are made and if it is realized that a test is the product of such extended and painstaking work.

CHAPTER FOURTEEN

The Testing Program

1. SELECTION OF A PROJECT

IN making plans for the use of tests, it is first of all important to formulate clearly just what it is desired to do. As has been said repeatedly, testing simply for the sake of testing is most unfortunate. First, there should always be the formulation of a very definite problem, in the solution of which the tests are to aid. Tests and procedure should then be determined with reference to this problem. And, finally, the value of the tests as an aid in dealing with this problem should be determined. If the matter is gone at in this way, much plundering about will be saved and a definite conclusion with regard to the nature of the service which the tests have rendered will be possible.

Projects involving tests of mental ability. In the use of tests of mental ability, a teacher may wish simply to obtain assistance in dividing her class into sections on the basis of ability. A principal may wish to determine whether there are in his school any children who are markedly misplaced as to grade. A superintendent may wish to compare the "pupil material" in the various schools. All these projects are admirable; and a "mental survey" will yield material for dealing with all three problems. In any extensive use of group tests of mental ability, data bearing on these problems and more will be gathered. Before attempting such a survey, there should be a careful formulation of these problems, and something planned with regard to action which may

be based on the test results and methods by which the usefulness of the tests may be determined. No one should ever test blindly, hoping that, somehow, some good may result. Instead, the person in charge should have clearly in mind just *what* good he hopes will result; and he owes it to himself, and to those who help in the work, not to leave the data without determining whether or not this good resulted. He should, therefore, so formulate the project that the success of the testing can be readily determined.

Projects involving tests in the school subjects. Projects which may be undertaken in using tests in the school subjects are even more numerous. Perhaps a teacher feels that the teacher who had her children the year before in the previous grade was weak in her teaching of arithmetic, and she wishes to verify this impression as a basis for assigning extra time to work in the fundamentals. An hour or so spent in use of the Courtis Tests will inform her in the matter, and the use of diagnostic tests will tell her just what she should do to improve matters. The supervisor may have devised certain special methods for teaching reading, and she wishes to determine their effectiveness. Again, tests will give her evidence of a definiteness which could not be obtained otherwise. Or, a superintendent fears that not enough time is being devoted to spelling. Once more tests will help him when nothing else can.

Use of a test with reference to a particular problem should not mean, of course, that any bearing which the results may have upon other problems should be neglected. Use of tests without any definite object in view is unfortunate, but so is neglect of the richness of implication

which test results usually have. It should be remembered that any use of tests involves an investment of both time and energy, and one should always aim to obtain the greatest possible return on such an investment. The tabulations should always be studied carefully. Along with the statement of the major objective it is often well to list possible further values which might be expected. Thus the supervisor may use a reading test primarily to determine the efficiency of her pet method of instruction in reading. The resulting data will also be useful, however, in the comparing of the work of different teachers, will yield some diagnostic information with regard to the "rate" and the "comprehension" of reading on the part of the children tested, and will be of value to each teacher in pointing out those children in her class who need special help.

Always, then, in planning for use of tests there should be formulation of a definite problem, to the solution of which the tests are to contribute; and there should be a further effort to obtain all possible values from the test results. In planning any testing program, such very definite formulation is essential if the tests are to be as valuable as they should be, and if any intelligent opinion as to the value of the tests is to be reached.

2. SELECTION OF TESTS

A recently published bibliography of tests includes a total of 278 titles! Evidently it is now possible for one to make a choice of tests; no longer is one limited to a single first crude instrument in any of the major school subjects. So, once a project has been formulated, the next question is as to what tests shall be used in dealing with the problem chosen.

As a matter of fact, consideration of the tests available may modify or even determine one's choice of problem. Though a large number of tests is now available, it is nevertheless true that tests in certain fields are still, because of poor test form, so expensive or so laborious to use that a teacher or superintendent may well hesitate before undertaking a project in that field.

Consideration of the general nature of the test. In considering various tests which might conceivably be used in dealing with the problem chosen, it is first necessary, obviously, for the superintendent or principal to become familiar with the general nature of these tests. The pupils' blanks, direction sheets, scoring keys, record sheets, should all be carefully studied. The nature of the questions or items in a scale should be considered, and the way in which these items were selected. The adequacy of the norms should also be looked into. In short, the principal or other person in charge of the work should, in considering each test, make a detailed study of the test materials and the way in which the test was constructed.

Consideration of the practicability of the test. Once the principal or superintendent has clearly in mind the general nature of a test, he is in a position to consider its practicability. Perhaps such careful consideration of practicability is the matter most often neglected in planning a testing program. Over and over again it happens that teachers are rendered hostile to tests because they have been asked to use instruments which loaded their already full schedules to the breaking point. It was asserted in the first chapter that tests should *save* time. Unfortunately there are many tests now in use which require an expenditure of time and

effort all out of proportion to the value of the results yielded. Testing should not be one more imposition upon the teachers. If testing cannot be done with their coöperation, it should not be done at all. Superintendents and principals should very carefully consider the matter of time costs, and it should be their definite effort to choose only tests which are easily given and scored. Always, any testing program should be mapped out with careful consideration of the relation of the teachers to that program, and tests should be chosen which are useful to them and as little burdensome as may be.

It is also worth mentioning that in beginning the use of tests in a school or school system, it is particularly advisable to choose a test which is easy to use and which has simple and full directions. From such a test general principles and methods of statistical handling can most easily be learned — and the essential facts of method will emerge and not be confused with the many unessential details which some of the more complicated examinations involve. In introducing tests into a school system, one should be careful not to undertake too much, and not to create the impression that testing is a laborious and complicated matter. The future of testing in a system will depend quite largely upon the first impression created, and the nature of that impression will be conditioned chiefly by the nature of the first tests used. Especially in beginning test work is it very important *not* to choose a test of such poor form that a large amount of labor is required in the use of it.

Consideration of the use for which a test is intended in relation to the problem. In choosing tests for a project, it should be obvious that the superintendent must first familiarize himself with the general nature of the

tests under consideration; and in this examination of the test materials a decision is naturally also reached with regard to the practicability of a test as a means for dealing with the project in hand. However, all this is largely prefatory to the major consideration — the consideration as to whether or not the test is really just the type of test to yield exactly the information desired. That is, is the use for which the test was intended exactly the use which is contemplated in the project under consideration? It should be obvious that if a teacher wishes guidance in her instruction in arithmetic she should not choose the Courtis Series B; she should choose a diagnostic test. It should also be remembered that supervisory tests are intended primarily for the measurement of groups; it should, then, not be expected that results on individual pupils from these tests will be highly reliable. If a superintendent wishes to use a test of mental ability to gain a general idea of the "pupil material" of his schools, he should use a rapid survey examination designed for just such purposes; if he uses a more detailed examination intended to give a reliable diagnosis of individuals, he will have to spend a great deal more time on his project than he needs to.

It is impossible to be too explicit in the formulation of a project, or too exacting in one's demands that the test chosen be very specifically suited to deal with the problem in hand. Care should be taken to choose tests which are known to be of value for exactly the use intended, particularly if one is just starting in with test work. If one plans to use a group test as an aid in section division, it will be well to use a test which has been reported as useful for this purpose, and to use the test exactly in the way recom-

mended in the reports. If one wishes to improve the work in arithmetic as a result of the employment of a series of diagnostic tests, then one should adopt a series of established worth for just such purposes — and a series in connection with which there are definite data demonstrating such value and making clear the way such improvement may be gained.

One must also remember that special circumstances may qualify results in dealing with special types of cases. Thus in work with children from homes where English is not spoken, group tests of mental ability of the type now common for work in the upper grades must be employed with caution; a non-language or performance scale may be desirable.

Limits of applicability must also be kept in mind; a test for use in the grades would be of no use in high school, for instance. A handwriting test is hardly worth trying in the eighth grade. In fact, it is a fairly safe procedure to limit — for practical purposes — the use of tests in a given subject to those grades in which that subject is specifically *taught*, as in these grades only will there be a sufficient opportunity for the remedial instruction that should follow the use of tests.

It should be obvious that it is exceedingly important to choose carefully the tests one plans to use in a school or school system. Nevertheless, tests seem to be chosen frequently upon the recommendation of a friend without any real thought of their worth; and particularly without due consideration of the specific needs of the particular project to be dealt with. In the above brief section it has been urged that a test should be decided upon only after very careful consideration of (1) its general nature, (2) its practicability, and, most important, (3) its usefulness — its suitability for the purpose of solving the problem at hand. Such

careful thought will be more than repaid by the value of the results.

3. ORGANIZATION OF TEST WORK

In planning a testing program, one of the first perplexing questions is concerned with the organization and administration of the testing. Unquestionably the test should be bought by the school system; teachers should not be asked to pay for blanks. The question then arises as to whether the teachers should give and score the tests. No categorical answer can be given to this question; it depends upon the nature of the test, the nature of the problem, and the facilities for such work at the disposal of the system. It is now becoming common for a school system to have a "director of research" who gives all or part of his time to such work; in large systems there may be a definite bureau with a regularly constituted staff. When there is such a bureau, it is desirable that some one connected with it supervise the testing. This is desirable especially in the use of certain tests, because the giving is difficult; in extensive work involving school comparisons it may also be well to have some one from the central bureau do the testing, so that the personality of the tester may be the same for all schools to be compared. It is only fair that in using tests in the study of administrative or supervisory problems rather than teaching problems, the teacher should not be asked to do all the work.

The part of the teachers in the testing. However, in many systems there is no such bureau specializing in the use of tests; and where such a bureau does exist, it is rarely equipped to take care of all the work required

in an amount of testing sufficient to justify its existence. In other words, practically always some assistance from the teachers is necessary, and it is desirable. The tests cannot function as they should unless the teachers are interested in them and understand them; and this interest and understanding can hardly be attained without the participation of the teachers in some phase at least of the test work. The entire burden should not be upon them; neither should they have no share in the matter. The testing should be a coöperative project, in which both teaching and administrative staffs have their parts.

In general, it is desirable that tests should be given by the teachers. The most modern tests can be so handled; and if tests easy to give have been chosen, there should be no difficulty in this respect. Testing should always aid the teacher; and where the tests are to function primarily in assisting the teachers, the testing should be in their hands. Except in the special cases above mentioned, where the testing is primarily with reference to large administrative problems, the teacher will also feel much easier about the matter if the testing is largely carried on through her. Both giving and scoring should, under such circumstances, be done by the teacher. The scoring will not burden her greatly if the choice of tests has been wise, and the teacher will understand the meaning of the score much better if she has done the scoring than would be the case otherwise. Further, there will be no possibility of feeling on her part that the tests have been inaccurately or unfairly graded. Finally, if the teacher does the scoring she will learn much in going over the papers

about the performance of each child which does not appear in the final total score at all; she will discover many little details in the work of each child which she would not learn otherwise — will be able to use even the general tests, to a certain extent, for diagnostic purposes.

Once the tests have been given and scored, the next question is as to who shall tabulate and interpret the results; and again it may be said that the work should be, so far as possible, in the hands of the teachers. It is desirable that the teacher gain experience in such work, and it is vital that she should understand and be sympathetic with it. Only by having a real part in the work can such an attitude be developed. Clearly, under such circumstances, the teacher must not be asked simply to give and score tests — the drudgery of the whole business — and then turn the fascinating business of interpretation over completely to a director of research. Besides, the teacher can often help in the interpretation.

The contribution of the central office to the test work. Perhaps it is not quite clear from the above statements as to just what part a director of measurements (where there is such an individual), or the superintendent or the principal or other person in charge, may have in the whole undertaking.

Roughly, the central office should function in two ways in the testing. It should do as much as possible of the testing having to do with the larger administrative and supervisory problems; and in the second place the office should instruct and assist in that major part of the testing which has to do with instructional

problems. This notion that the director of measurements, or the superintendent or principal, should assist in work which is recognized by the teachers as beneficial rather than impose a type of work which the teachers resent as inquisitorial, is so important that the following chapter will be devoted largely to the topic.

It has been assumed in the above discussion that the administration takes the first move in connection with any testing project. Unfortunately this is not always the case. Sometimes the administration is not in a position, — financially, or for some other reason, to do this. Sometimes, unfortunately, a superintendent is not so progressive as some of his teachers; however, it is only fair to say that it is sometimes wisest to have any such method first introduced into a system informally through the action of individual teachers. When the teacher is doing the testing on her own responsibility, and simply as a matter of her own interest in such work, it is obvious that the above section does not apply. Of necessity the work is all in her hands. It should only be pointed out that she owes it to herself to choose tests which are easily given and scored, and tests which are not primarily supervisory instruments but which will help her in dealing with her instructional problems. If, as urged previously, she will clearly formulate her problem and then choose her test in the light of that problem, she will be saved many mistakes.

4. PLANNING THE TESTING PROGRAM

It has been suggested above that most testing will be done by teachers, with the advice and assistance of the central office. The way in which that advice and assistance should be given has been reserved for discussion in the following chapter. Certain details with regard to a testing program remain to be mentioned.

When the tests should be given. In the first place, there is a question as to when tests should be given. Obviously this will depend upon the nature of the tests, the problems to be dealt with, and the general circumstances. If one is using group tests of intelligence as a basis for aid in making section divisions, or in advising high school freshmen with regard to electives, it is clear that the testing should be done at the very beginning of the school year. If a teacher is using tests in the school subjects in order to find out just where she should begin her instruction with a new class, then also should the tests be given at the very start of the school year. If, on the other hand, the tests are to assist in the decision as to what children should be failed in a semester's work, or are to serve in appraising the efficiency of the teaching, then, as obviously, the testing should be done near the semester's end. It is also worth while noting that there is sometimes a day or so at the beginning of school when things have hardly settled down enough to make efficient instruction possible; tests may be given at such a time without interfering with instruction to an appreciable extent. Something of the same situation is true with regard to the last two or three days of school; there is a general stir and let-down in morale which interferes with instruction but affects testing much less. It is further to be mentioned that norms for some tests are stated in terms of testing either at the beginning or end of the year, or in the middle of the year; it is usually possible to allow for differences in the time of testing, but if it is convenient it is probably better to use the tests at the time indicated. The beginning or end of a

semester are, for many purposes — such as double promotions — the most advantageous times for testing. However, Binet examinations and diagnostic tests in the various school subjects may often be of service in the middle of the semester, or at any other time that the teacher feels the need of such assistance as these tests can give.

The number of tests to be given. The next question is as to the number of tests which should be used in a given year. No categorical answer can of course be given. In initiating a testing program it is better to undertake too little than too much; it must always be remembered that there are a great many things about testing that take time, which one hardly thinks of. The writers have known a school system to lose good teachers because the system "plunged" in testing and fairly worked the teachers to death. So the first caution is, in initiating a testing program, to "go slow" and not undertake too much, but to do well what is done, so that the teachers will realize its value. The thing to do is to choose a very concrete specific problem, and then see that one thing through in clean-cut fashion. Perhaps it may be advisable to begin with tests in only one school subject, or with tests of intelligence alone.

If more than one test is used, the tests should make a coherent program; indiscriminate and merely extensive testing is wasteful. It may be well to start with a program limited to one problem, such as the improvement of instruction in arithmetic. Thus one might begin with a general test in the fundamentals of arithmetic, follow this with a diagnostic test and with a campaign

for remedial instruction based on these diagnostic findings; and then use general tests again at the end of the semester in order to measure the improvement. Or, if one wishes to investigate the usefulness of diagnostic tests, one might put through the general tests in the fundamentals in several schools, then use the diagnostic tests in only a few of these schools, finishing up with another general survey at the end to show the relative gains made in schools having remedial instruction and in those not using the diagnostic tests. Or a systematic plan for readjustment, using group tests of mental ability, may be tried. The particular problem chosen is a relatively unimportant matter; the fundamental thing is that there should *be* a clearly defined problem, and that the project should be seen through to the end, with a proper publicity given to the results.

If it is desired that there should be testing in a number of subjects, it may be desirable to employ a combination examination that contains tests in several subjects. Such examinations will be found greatly to simplify procedure. There will be much less confusion, multiplicity of blanks and of methods; and the study of the results in their interrelations and the significance of the combined scores will be easy, where it is almost impossible otherwise.

A number of such examinations have appeared. Frequently they are simply combinations of standardized tests which have already appeared separately but are printed together in a single booklet. Some few of them present altogether new tests especially designed for such use. Some of these examinations cover a number of grades. Others with more detailed material are applicable to one grade only. But whatever their exact nature, they are

intended to simplify the work involved in a comprehensive survey.

The Pintner Educational Survey contains tests in all the fundamental subjects of the grammar school, and is applicable from the third grade through the eighth. The writers' Scales of Attainment, Nos. 1, 2, and 3, are limited to Grades 2, 8, and 3, respectively; they are designed to measure the "essential achievement" of the pupils in these grades. The Illinois Examination contains tests in reading and arithmetic, and also a complete intelligence examination; it is more elaborate than the other combination scales mentioned. It permits, however, of comparisons between a child's intelligence and his work in school in these two subjects. There are forms applicable from Grades 3–8.

As the testing in school systems develops, use of a greater variety of tests will be found worth while; and it will be discovered that with the increase in the number of tests used there is an even greater increase in their usefulness, since each added bit of data throws light on all the other results, and is in turn made more understandable by these other findings. One might almost say that as the number of tests increased arithmetically, the values yielded increased geometrically.

Summary. In arranging for test work, then, four steps are advisable. First, there should be a very careful and specific formulation of the problem, or problems, with which it is desired to deal. Next, tests should be selected which will deal efficiently and directly with this problem, and which will be practicable to use. Following this should come a careful organization of the work and a decision concerning the part of each person concerned. Finally, plans should

be made as to the time of testing, the number of tests to be used, relation of the particular work in hand to previous testing and to future work, and so on. It is hardly possible to give too careful attention to this working out of the program. It is because of a lack of program that so much testing is "waste" work.

CHAPTER FIFTEEN

Making the Testing Program Worth While

1. Enlisting the coöperation of the teachers

IT has been stated in the previous chapter that for the most part tests will be given and scored and first interpretations made by the teachers. It should go almost without saying that such extensive work should not be imposed upon them by mandate from the office. Instead, every effort should be made to secure the intelligent and enthusiastic coöperation of the teaching force. Much of the opposition shown by teachers to the use of tests arises from the failure of some person in authority to make clear to them *why* the tests are being given, what values may result, and just what their part in the undertaking is to be. The testing may well be initiated at a teachers' meeting by a careful presentation of the proposed plan for test work, with full explanation of the way in which it is hoped the tests will benefit the teachers. It may even be well to make the first formulation of the problem a matter for consultation with the most able teachers in the school or system. At the teachers' meeting this problem should be discussed, and the test or tests which it is proposed to use should be shown and explained; sometimes it is well to acquaint the teachers with the nature of the test by giving it to them informally and having them score their own papers.

At this first meeting nothing more may be done than to discuss the general nature of the problem and of the particular tests which are to be used. Special

emphasis should be put upon the service which it is hoped the tests will render, and every effort should be made to have the teachers feel that they are being consulted in the matter, that their opinions are being taken into account, and that the results of the tests will be of distinct value to them. In closing this first meeting the teachers should be given sample blanks and direction sheets, and be asked to study these materials over and be ready at the next meeting with any questions. At the second meeting procedure in giving and scoring should be gone over in sufficient detail to make certain that every one understands exactly what is to be done. Then, within the next few days, the tests should be given. After sufficient time has elapsed for the scoring of the papers there should be a third meeting for interpretation and discussion of results. This last meeting is absolutely essential for the progress of testing in a system — and it is the meeting most frequently omitted. The writers have in mind surveys which were made at the order of the central office, in which the tests were given and scored by the teachers, the results reported back to the central office by the principals of each school — and the whole project never heard from again. The school system, in such cases, gets some ordinary clerical help out of the teachers for nothing. Such a situation should never occur; there should *always* be a getting together of the teachers concerned with the testing, and a joint consideration of results from the various classes.

It should be obvious — but unfortunately seems not to be so — that if the testing program is to be worth while it must be carefully organized and put through in a business-

like manner. This means that blanks must be ready and at the proper schools when they are needed, teachers prepared to give the tests, the time of testing arranged, and so on. All this would seem to be self-evident. Yet on one occasion a big project in a large school system was so clumsily planned that in certain schools the test blanks and directions were not received until a few minutes before the time set for the examination; some of the teachers had never seen the test previous to this time. In other schools the materials had been carefully discussed at a teachers' meeting, and the teachers had had a week or two in which to familiarize themselves with the directions. The survey was made primarily for the purpose of comparing schools. It would seem painfully obvious that no such comparisons could be made; the results would certainly differ considerably because of the differing degrees of preparedness on the part of the teachers, regardless of any differences of ability among the pupils. Such obvious errors as keeping one room in during recess when every one else is outside, or keeping the whole school inside over the recess period, or giving a test the last thing before closing time, or holding the school after time for dismissal, should certainly be avoided.

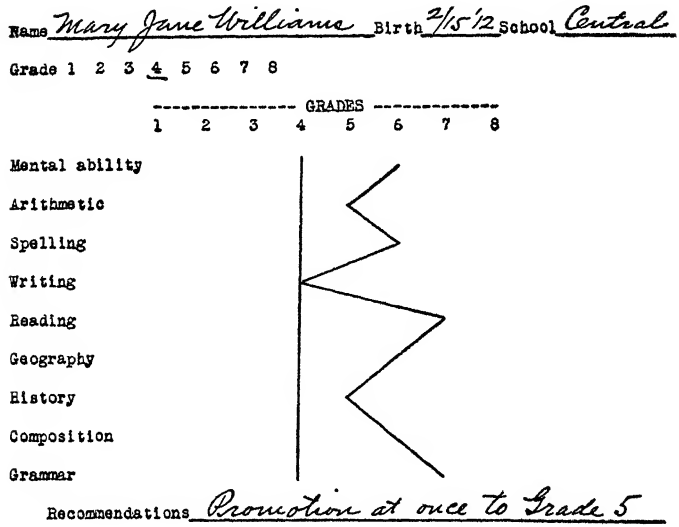
Throughout, every effort should be made to present the tests as instruments to aid the teachers. It is very unfortunate that since tests in the school subjects were first used in connection with supervision, there is the impression among many teachers that tests are used primarily to investigate the efficiency of their teaching. Everything should be done to prevent the arising of the notion that test results are to be used as a basis for "firing" a teacher if her class fails to do well on the tests. Very rarely is it desirable to use tests in this way; and in any case it is only one of the dozens of possible uses and should not be allowed to dominate the situa-

For example, we will suppose that a group test of mental ability has been given, and that Mary, who has been doing poor school work, tested at the top of her class in ability. One should not accept this at its face value without investigation; the situation must be studied closely. As in a case of which the writers knew, it may be found that the girl had recently come from another school in which she made excellent marks but that she has somehow failed to "get along" well in her new environment, has become morose, and has allowed her work to "slump." Her teachers feel that she is capable of much better work than she has been doing. Thus the full value of the test score in such cases appears only when the score is used along with other information; and when the test is used along with marks, teachers' estimates, and other kinds of information, all these sources are made more valuable than they otherwise would have been. However, merely to understand Mary's situation is not enough; something must be *done about it.* The information is of no value whatever unless some action is taken. In the particular case above referred to the teachers agreed to make a special effort to interest the girl in her school work again. They also planned to work through two or three of the most tactful of the girl's classmates to draw her into the good times of the school and put her on good terms with the other children, as she had been in her former school. The result of all this was a regaining of the good marks she had always before received, and it should be remembered that the test score — which was so out of harmony with the school work — was the starting point of the inquiry and of the resulting adjustment.

So every effort must be made to interpret intelligently the test findings, every effort must be made to cause the test results to function in school work; and it is in the interpretation and application of the test findings that a director of research can be of great assistance to teachers, because of his greater experience with test results.

It should be the attempt of the teachers, or other investigators, to express test results in a form which will be useful and will show clearly the relation of the test scores to other information. Hence the value of grade norms, and of graphs and charts exhibiting in a vivid way the comparative standing of different children, their standing in addition as compared with subtraction, in arithmetic as compared with geography, or their increase in ability in arithmetic consequent upon use of practice exercises. The person in charge of the testing should make a special study of such devices. Sometimes it is worth while to have a special card for each child on which test results are summarized, so that a glance will reveal the comparative standing of the child. Such a card appears on the following page.

On this card there are spaces for scores in mental ability and in eight fundamental subjects of the grammar school. The vertical line represents the grade placement of the child at the time of testing — in the fourth grade; the broken line indicates her standing on the various tests. She tests above the norm for the sixth grade in mental ability, geography, spelling, and composition; above the norm for the seventh grade in reading and grammar; and above the norm for the fifth grade in arithmetic and history. She tests in the grade in which she is placed in only one subject — handwriting — which is certainly not a sufficiently



important subject to hold her in that grade. Such a child could certainly be advanced at once to the fifth grade and, if her teacher thinks it advisable as far as her health is concerned, to the sixth. Her test results show that she could probably do the work of the sixth grade if she could be given a little extra work in arithmetic; the history she would undoubtedly make up in the course of a semester of work with sixth-grade pupils. It makes the facts a little more clear to put the mental ability standing in red, as that is perhaps more fundamental than the other scores in the particular subjects. Such a card summarizes the test results and gives a picture of the total standing of a child. If teachers and others would make use of such summary cards as this, there would be less "lost motion" in testing.

A variety of such devices is possible, and they will be found distinctly useful. Once the teachers have

become familiar with test methods, and have learned ways of interpretation, they will constantly be finding ways of using the test findings in dealing with their problems. Such ways have been suggested in previous chapters. It remains only to point out that a satisfactory system of records, making these test findings readily accessible, will cause the teachers to refer back to the tests more and more in dealing with problems as they arise. A cumulative record of test results on all children of a school is of great value to the teachers and well worth the little extra time needed for recording the data in an orderly and systematic fashion; if care be taken not to allow this cumulative record to become unwieldy, it can be of great service. For instance, a fifth-grade teacher finds that James is having trouble with arithmetic. Consultation of James's record shows that on a set of diagnostic tests the year before he had trouble with all forms of division and with "carrying" in addition. A little special help in the matter puts James's work in good shape again. Or, Molly is doing poor work in geography, and her teacher is doubtful as to whether she should be passed or not. A study of Molly's test record shows that her score in geography is three half-grades below the grade she is in, and that her score in mental ability is even lower than that. It becomes clear, then, that if grade standards are to mean anything at all, and if children should ever fail, Molly should fail in geography; there is no one else in the class who tests so poorly, and there are pupils in classes below her own who test markedly above her. If Molly's parents object, there is this striking test record to show them as evidence that the failing mark

is not merely teacher's bias. Or, Henry is so quick and accurate in arithmetic that it hardly seems worth while, for the present, to drill him further in the fundamentals; instead, the teacher wishes him to work on spelling, in which he seems to be weak. Consultation of the test record shows Henry to be far above his class in arithmetic, but below in spelling, and gives the teacher full warrant for excusing him from his drill.

3. VERIFYING THE USEFULNESS OF TESTS

It is of fundamental importance that the test results be actually used in the school, and as a result of such use some definite conclusion should be reached as to their value. Those tests which are not of practical worth will then stand forth clearly as valueless. The comparative usefulness of different types of test may be worked out, and the best methods for employing tests for dealing with educational problems may be determined. Such verification, or checking up, on the test results is of essential importance; and, as was said at the beginning of the previous chapter, it is well at the start to plan the project so that definite checking up may be possible.

Verifying results with tests of mental ability. If, for instance, a teacher has used a test of mental ability and has made up section divisions on this basis, she should certainly notice whether or not the divisions thus obtained are more homogeneous in mental ability or brightness than similar divisions made in previous years on the basis of her own judgment. She may compare the marks she assigns to the class and see if they are more closely grouped than usual. That is, she may

find that the year before, when she made up her section divisions on the basis of brightness as judged by her according to her own observation of the children, there were five children in her "slow" section who received marks distinctly above average, while in the "slow" section as chosen on the basis of brightness according to the test results there were no children who received good marks. In other words, the test results have helped that teacher by grouping the children so that the work done by those in each group was of about the same level. Or, if the teacher has used test results to back up her judgment in giving a child a double promotion, she should keep track of the child's progress and see whether or not the move was a wise one.

If a principal has used results of tests of mental ability as a basis for reclassifying his grades, he should carefully check up, at intervals of a month or so, to find out how many of the total number of shifts have worked out successfully. Thus, if he has changed the grade placement of 24 children, he should be able to state, at the end of the next semester, what proportion of the 24 changes are to be permanent; he should find out from the teachers the level of work done by these children, and get from them an estimate of the correctness of the present placement. If he finds that 85 per cent of those moved are to remain in their new grade, he may conclude that the test results have been of real value to him. If a superintendent has used mental tests to make a survey of his system, he should check up any differences he may find between various schools by a general survey of the neighborhoods which the various schools serve; and if the test results have helped him to

understand the problems of the various schools and neighborhoods better than he did before, they have made a real contribution.

Verifying results with tests of achievement. Similar conditions hold for the verification of test results in the school subjects. If a teacher has given a diagnostic test in the fundamentals of arithmetic (preceded by a general test), she should follow up any remedial teaching with another general test, that she may find out whether or not her remedial instruction — undertaken on the basis of test results — has really resulted in a better grasp of arithmetic. However, she may not wish to verify test results in terms of other test results. She may then use her own observation as to the progress of children, to whom remedial teaching of a very specific nature has been given, as compared with the progress of other children whom she has taught without the aid of tests. If a principal finds that the pupils of his school seem to be below average in the "comprehension" of what they read, according to a test of reading, he should call together his teachers and try to find out from their reports of their own observation whether or not this is so; and, if he uses some special device for improving the degree of comprehension, he should check up and find out whether or not this practice has resulted in a real gain.

Some principals of junior high schools have made rather extensive use of test results to classify children according to their standing in various subjects. That is, they have placed each child in each subject in the grade indicated by his score in that subject. Thus, John may be in 7B in arithmetic, 8A in history, entirely

through with geography, taking something in high school — and so on. If any such extensive reclassification is attempted, or even if only a few children are changed about according to the test results, careful record should be kept of the permanency of the changes; and the person in charge of the experiment should be able to state the percentage of accuracy yielded by the tests in placing children in their proper grade by subject.

In general, then, there must be very careful verification of test results, so that the users may be able to say definitely just *what* good they have gained from the tests — just *how* valuable the tests have been. One frequently meets people who are enthusiastic advocates of the test method but who are entirely unable to give a single piece of concrete evidence to back up their faith in the method. The teacher or principal who is convinced of the value of tests can do nothing more helpful in furthering their use than to show in definite figures just what services tests have rendered in his class or school. This verifying of the value of the tests is, of necessity, the last step in the testing program, but it is an exceedingly important step.

4. THE LONG-TIME TESTING PROGRAM

Once the teachers are properly initiated into the use of tests, they will soon learn to handle and apply the tests, and verify their value for themselves; and testing will no longer appear a perfunctory bit of procedure, a part of the red tape which seems a necessary evil in a school system. Instead, tests will come to have an essential place in the work of every teacher, and the teachers will wish not less testing, but more. They will wish

measurement in *all* the school subjects — will wish to compare the performance of each child in the various subjects, to make detailed comparisons between different classes and schools. They will find that they need measurement of both ability and school work, that they need to study school achievement with relation to ability.

As the test work grows it will be realized that a carefully coördinated program having reference not merely to present needs, but taking account also of the needs of the next year and of the year after that, is required. The data yielded by tests should be cumulative. There should be a record system which will make these cumulative records readily available. Testing should be regularly done at certain established times in the school year. Children will come to follow their work as their progress is revealed by the tests, and will have their work largely motivated in this way. The tests will be a vital part of educational procedure. It is unwise for a school system to initiate a test program with any efforts at such extensive employment of the method; but an even greater mistake is made when testing is treated as an incidental and perfunctory matter and the possibilities of such development are not appreciated.

APPENDIX A

FINDING THE MEDIAN FOR LARGE DISTRIBUTIONS

In the chapter on statistics the median has been defined as the middle case. When thus defined, the resulting figure is sufficiently accurate for study of small groups, as classes, grades, or small schools. However, when large numbers of cases are involved, some method permitting a more refined treatment of the data, with calculations of the median to one or more decimal points, becomes desirable. In this appendix such a method will be described. The method is a somewhat arbitrary, though natural, development from the "common-sense" concept presented in the earlier chapter. For the purpose of this appendix the median will be redefined as the middle point of the distribution—not the middle case. Directions for locating this middle point in (*a*) an ungrouped distribution and (*b*) a grouped distribution may be briefly described.

(*a*) The distribution below represents the scores made by 2346 school children in the fourth grade of several cities. It is ungrouped, the scores running from 1–16.

Score	No. Cases
16	48
15	57
14	102
13	155
12	223
11	276
10	294
9	279
8	246
7	188
6	156
5	101
4	88
3	57
2	46
1	30
	2346 Total No. of Cases

The method of finding the median for such a distribution according to the rough method given in the first section of the book is inadequate where it is desired to make a comparison of medians which are nearly equal in value. If the reader will go over the distribution according to the method there suggested, he will see that the median is nearer 10 than 9. It will also be appreciated that with this large number of cases fractional values might well be significant — as in comparing two school systems. A consistent and direct rule for finding the middle point of such a distribution is as follows:

(1) Divide the total number of cases by 2. (2) Count up from the bottom of the distribution to that interval whose cases, if added in, will make the sum greater than half the number of cases; the median is somewhere within this interval. (3) To locate it exactly, divide the number of cases still needed to reach the median by the number of cases in this interval. (4) Multiply the quotient by the "amount" of the interval. (5) Add the product on to the lower limit of the interval.[1]

This rule may seem difficult, but may be made clear by a concrete example of its application. The calculation of the median for the above distribution may be made as follows:

(1) Divide the total number of cases by 2.

$$2\overline{)2346}$$

1173 — half the No. of cases

[1] In the above rule the term "interval" refers to the units of the distribution. Thus, in the example above, each number in the column to the left represents an interval. In the second example, to follow, each pair of numbers represents an interval. The "amount of the interval" refers to the grouping. If the distribution is ungrouped, as in the above example, the amount of the interval is 1. In the example to follow the distribution is grouped in twos, and the amount of the interval is 2; if the grouping were by fives, the amount of the interval would be 5, and so on for any other groupings. The "lower limit" of the interval means simply the beginning of the interval. Thus in the illustration above, the first interval may be considered to extend from 1.00 to 1.99; the lower limit would thus be 1.00. In the next illustration the first interval extends from 1.00 to 2.99; but the lower limit remains 1.00.

(2) Count up from the bottom of the distribution to that interval whose cases, if added in, will make the sum greater than half the number of cases. The median lies within this interval.

30 plus 46 (76), plus 57 (133), plus 88 (221), plus 101 (322), plus 156 (478), plus 188 (666), plus 246 (912). If the cases in the next interval are added in, the sum will become 1191, which is greater than half the number of cases. Since all the cases in the 8th interval have been added, and the cases in the 9th interval make the sum too large, the median must lie in the 9th interval.

(3) Divide the number of cases still needed to reach the median by the number in the interval in which the median lies.

1173 — half the No. of cases
912 — sum through 8th interval
261 — No. still needed to reach the median
279 — No. of cases in 9th interval
.93 — quotient

279) 261. No. still needed

(4) Multiply the quotient by the amount of the interval. (This step may be omitted in dealing with ungrouped distributions, of course.)

Since the distribution is ungrouped, the amount is 1.
$.93 \times 1 = .93$

(5) Add the product to the lower limit of the interval.

The lower limit of this interval is 9.00. Therefore,
$9.00 + .93 = 9.93$.

The median is 9.93.

(*b*) Below is presented the same distribution grouped by twos. The procedure in finding the median is exactly the same, and will be somewhat abbreviated. Step (4) should be specially noticed.

Score	No. Cases
15–16	105
13–14	257
11–12	499
9–10	573
7– 8	434
5– 6	257
3– 4	145
1– 2	76
	2346 Total No. of Cases

(1) Half of the total number of cases 1173

(2) Sum of the cases in the first four intervals 912

(3) Cases still needed to reach the median 261

Number still needed divided by the number of cases in the 9–10 interval . . . . . 573) 261. (.45

(4) Multiply quotient by the amount of the interval . . .45 × 2 = .90
Since the distribution is grouped by twos, the amount of the interval is 2.

(5) Add the product to the lower limit of the interval 9.00 is the lower limit of the interval 9–10
9.00 + .90 = 9.90

The median is 9.90.

It will be noticed that the two medians, calculated from the grouped and from the ungrouped distributions, differ slightly; such slight differences are to be expected with different groupings. These slight variations are negligible. In general, the median calculated from an ungrouped distribution may be considered a bit more reliable.

The above method may seem at first somewhat elaborate; but a little practice will make the steps clear, and the entire process will be found easy.

APPENDIX B

TESTS AND THE DIAGNOSIS OF FEEBLE-MINDEDNESS

In the text relatively little has been said about the use of tests of mental ability to find feeble-minded children. Such discussion has been omitted partly because the subject is highly complex, partly because the teacher is hardly competent to make such a diagnosis, partly because in the past too much emphasis has been put upon use of tests in just this connection and it was desired that attention should be diverted to other and more profitable uses. However, something more should be said with regard to these matters — if only to make clear this general point of view.

In the first place, it should be appreciated that feeble-mindedness is a distinctly complex condition — that, in fact, there are various kinds and degrees of feeble-mindedness, so that no simple rules for the detection of such conditions can well be given. Feeble-mindedness is not simply inability to pass certain tests. Feeble-mindedness is a profound and subtle mental inadequacy, existing from birth or from an early age, and frequently accompanied by physical, mental, and moral abnormalities. Only by careful study covering all these various points, and going into the life history of the individual and into the family history as well, can an adequate diagnosis be made. No child should ever be diagnosed as feeble-minded simply on the basis of a test score; all these other factors must be considered. The feeble-minded child shows his mental defect by gross inability to learn in school, by inability to get along with other children, by a tendency to play with children much younger than himself, by retardation of physical growth, and frequently by outbursts of temper, delinquencies, and general lack of moral restraints. The adult feeble-minded show these same asocial characteristics, together with economic incapacity — gross inability to earn a living.

In dealing with such cases the tests of mental ability are simply one factor bringing one to the final conclusion, but they constitute a very important factor. Here, as elsewhere, they bring out in very definite and explicit fashion facts which will otherwise be largely qualified by the prejudices of those concerned, or obscured by special circumstances. A child may do poorly in school, not because he is feeble-minded, but because he is flighty and temperamental and will not apply himself. He may have become delinquent because of a lack of home care, or because of nervous disease, without being mentally defective. In such cases the teacher may think the child feeble-minded. The tests will show that he is not. In other instances a feeble-minded child may be considered only a trifle slow by the teacher because he is so retarded as to be in a grade where the children are much younger than he, and because he is docile and does not attract attention to himself, as does the "bad" bright child. Here again the tests, with their impersonal procedure and carefully determined age norms, will reveal the truth of the matter.

So the tests are a great assistance in determining definitely whether a child is mentally subnormal or not. The question at once arises as to how poor a showing on the tests may be taken to suggest mental deficiency. In intensive study of individual cases (such intensive study as there should always be in considering any child who is suspected of feeble-mindedness) the Binet Scale should almost always be used. It may be said that, roughly, an IQ below .70 is suggestive of mental defect. This IQ must not be taken to *make* the diagnosis; but if the other facts regarding a child are confirmatory of such a conclusion, it may then be said to render this conclusion highly probable.

It must be appreciated, however, that there are no hard and fast lines separating "the feeble-minded" from the normal population. There are many more people who are simply

dull than there are feeble-minded. There is a large group of individuals who may be considered "subnormal" who are not dull enough to be considered feeble-minded; and just how subnormal a person must be before he can be considered feeble-minded can hardly be definitely stated. In short, there are all degrees of ability from extreme idiocy to genius, and there is no gap or break anywhere in the series. The feeble-minded are not a separate group by themselves; they are simply the lowest 2 or 3 per cent of the total population in ability. An IQ of .70 or below means an ability so poor as to be found among only about 2 per cent of school children.

It should be repeated again, however, that a low score on a test should not be taken by itself to warrant a diagnosis. The writers have had students give a child a Binet examination, taking less than half an hour, and forthwith hand in a diagnosis of "Feeble-minded." Such rash procedure is always most severely condemned. If the question is concerned with recommending a child for a special class, there should be careful consideration of his school record to date, his present work in school, his interests and character, the accomplishment of his brothers or sisters, the home and family conditions, as well as his test score. If all these factors suggest mental defect, a teacher may make a provisional diagnosis of feeble-mindedness, as far as her purposes are concerned. However, if the diagnosis is concerned with possible commitment to an institution for the feeble-minded, the whole matter is much more serious and a teacher should not presume to come to any diagnosis. The child should be examined by both a clinical psychologist and a physician, and the teacher should not confuse her diagnosis of pedagogical inadequacy with the final diagnosis which may be rendered by these experts.

It must be emphasized again that the teacher should devote herself to the consideration of the specially gifted children

rather than the defective children. Both the teacher and the school are better equipped to deal with the abnormally bright child than they are with the abnormally stupid, whose diagnosis and disposition should be left in the hands of experts.

APPENDIX C

SUGGESTIONS FOR FURTHER STUDY

THE present brief manual has been intended as an introduction to the use of tests. If it fulfills its purpose, it will cause the reader to desire further and continued contact with test work. So the final question with which the book will deal is the question as to how the busy school man may best keep informed with regard to progress in this field. Three suggestions will be offered:

(1) *Subscribe to at least one of the educational journals publishing material with regard to tests and their uses.* The following journals are recommended:

Journal of Educational Research: Public School Publishing Company, Bloomington, Illinois.

Elementary School Journal or *School Review* (according as the reader is interested primarily in elementary or secondary schools): Department of Education, University of Chicago, Chicago, Illinois.

Journal of Educational Psychology: Warwick and York, 10 East Centre Street, Baltimore, Maryland.

Teachers College Record: Teachers College, Columbia University, New York City.

Journal of Educational Method: World Book Company, Yonkers-on-Hudson, New York.

School and Society: The Science Press, Lancaster, Pennsylvania.

Journal of Applied Psychology: Florence Chandler, Publisher, Clark University, Worcester, Massachusetts.

The writers would suggest careful study of several issues of all these journals, with selection of the one or two which most fully meet the needs and interest of the reader.

(2) *Get in touch with the College of Education of your state university, or some similar reliable source of information regard-*

ing progress in the development and use of tests. During the past few years a number of bureaus have been established expressly for the purpose of furnishing school men with such information. The Bureau of Educational Research at Ohio State University (Columbus, Ohio), the Bureau of Educational Research at the University of Illinois (Urbana, Illinois), the Bureau of Educational Reference and Research at the University of Michigan (Ann Arbor, Michigan), the Bureau of Educational Service, Teachers College (New York City), the Bureau of Coöperative Research, Indiana University (Bloomington, Indiana), and the Bureau of Educational Measurements and Standards, Kansas State Normal School (Emporia, Kansas), all deserve mention. The World Book Company, Yonkers-on-Hudson, New York, has recently established a similar department for the service of its patrons. The progressive school man should ask to be put upon the permanent mailing list of one or more of these bureaus, and he should feel free to write the bureau, or the local university or normal school, with regard to his problems. From such sources he may obtain information with regard to the best tests to use for any project he may wish to undertake.[1] In certain instances he may be able to procure samples of tests and test materials from the same sources; otherwise information will be given as to where the samples may be secured.[2] The bureau, or the college of education, will also

[1] It is hardly wise for the average superintendent, principal, or teacher to attempt the selection of tests wholly on the basis of his own judgment. New tests are constantly appearing; since writing the body of this manual admirable tests have appeared in several fields, which will doubtless supersede some of the tests which the writers have described. Contact with one of these bureaus, or some other reliable source of information, will best keep the school man in touch with progress in test-building.

[2] It need hardly be said that study of sample materials should always precede adoption of a test for use in a school, and that such study is one of the best ways of becoming familiar with test work. In fact, a superintendent will do well to obtain samples of tests of the types in which he is most interested from the various publishing houses, and keep these on file in his office

be best able to advise as to further intensive reading along any special line in which the school man may be interested. Here again developments are too rapid to make possible any satisfactory bibliography in this volume; books dealing with the problems of feeble-mindedness, with the problems of measurement and method in the various school subjects, which were standard a few years ago are now out of date. Contact with the latest literature can best be secured through some such agency — and through a careful following of the journals and reviews.

(3) *As soon as possible obtain some first-hand experience in work with tests.* As with any line of work, the values and the problems of educational measurement can be adequately understood only when one has some background of actual experience in the field. A principal or teacher should not feel that he should take a university course in "testing," or wade through a dozen books on the subject, before he is prepared to use such instruments. The most modern tests are carefully arranged so that they can be given by the average teacher, without any special training along these lines; and no special equipment is necessary.[1] If due regard is paid to the elementary matters of technique and interpretation mentioned in this manual, there is no reason why test work

---

for study and reference; and if he is on the mailing list of these publishers he will receive notification of the issuing of new tests. The best-known publishers in this field are the World Book Company (Yonkers-on-Hudson, New York), the Teachers College Bureau of Publications (Columbia University, New York City), and the Public School Publishing Company (Bloomington, Illinois).

[1] However, a supply of quarter-inch cross-hatch paper, on which to make tabulations beyond those possible on the record sheet usually supplied with tests, will be convenient; and a slide rule (a fairly good slide rule can be bought for $3.00) will save much labor in calculating per cents, and in simple multiplications and divisions. Study of such a book as Carter Alexander's *School Statistics and Publicity* (Silver, Burdett & Co., New York, 1919) will be of much assistance to a superintendent or other person in charge of a testing program.

may not be at once begun, with the expectation that such work will prove both interesting and highly profitable. Study of results from one's own school or school system will be found the most profitable study of all, as regards both the tests and the school.

APPENDIX D

REFERENCES FOR FURTHER READING

I. THE following brief list of books may serve as a first guide for those desiring further reading in the field of testing. It should be understood that this list does not pretend to be all-inclusive; it is presented with the idea of giving the reader some clues to further reading.

A. Books Describing Tests, Methods, etc.

McCALL, W. A. "How to Measure in Education." The Macmillan Company, New York; 1922. 416 pages.

MONROE, W. S. "Measuring the Results of Teaching." Houghton Mifflin Company, Boston; 1918. 297 pages.

MONROE, W. S., DE VOSS, J. C., and KELLY, F. J. "Educational Tests and Measurements." Houghton Mifflin Company, Boston; 1917. 309 pages.

PINTNER, R., and PATTERSON, D. G. "A Scale of Performance Tests." D. Appleton & Co., New York; 1917. 218 pages. (Primarily a manual for use of the tests.)

TERMAN, L. M. "The Measurement of Intelligence." Houghton Mifflin Company, Boston; 1917. 367 pages. (Essentially a handbook for use of the Stanford Revision of the Binet Scale.)

WILSON, G. M., and HOKE, J. K. "How to Measure." The Macmillan Company, New York; 1921. 285 pages.

YOAKUM, C. S., and YERKES, R. M. "Army Mental Tests." Henry Holt & Co., New York; 1920. 303 pages.

B. Books Dealing with Statistical Methods

ALEXANDER, C. "School Statistics and Publicity." Silver, Burdett & Co., Boston; 1918. 332 pages.

KING, W. I. "Elements of Statistical Methods." The Macmillan Company, New York; 1912. 250 pages.

RUGG, H. O. "Statistical Methods Applied to Education." Houghton Mifflin Company, Boston; 1918. 410 pages.

C. Books Showing Use of Tests in Dealing with Educational Problems

School surveys are included in the list below because they offer very concrete material concerning the uses made of tests. A reading of such concrete

material is often more beneficial to the beginner than an attempt to absorb too much concerning general principles. Typical surveys of school systems of various sizes and types have been chosen.

COURTIS, S. A. "The Gary Public Schools. Measurement of Classroom Products." General Education Board, New York; 1919. 532 pages.

Cubberley, E. P., et al. "School Organization and Administration. A Concrete Study Based on the Salt Lake City School Survey." World Book Company, Yonkers-on-Hudson, New York; 1916. 346 pages.

JUDD, C. H. "Measuring the Work of the Public Schools" (based on the Cleveland Survey). Cleveland Foundation Survey Report, Cleveland, Ohio; 1916. 296 pages.

JUDD, C. H. (Director), et al. "Survey of the St. Louis Public Schools." Part One, Organization and Administration, 276 pages; Part Two, The Work of the Schools, 359 pages; Part Three, Finances, 246 pages. World Book Company, Yonkers-on-Hudson, New York; 1918.

JUDD, C. H., and SMITH, H. L. "Plans for Organizing School Surveys, with a Summary of Typical School Surveys." The Thirteenth Yearbook of the Society for the Study of Education, Part II. Public School Publishing Company, Bloomington, Illinois; 1914. 85 pages.

"Public Schools of Rifle, Colorado." Sentinel Press, Grand Junction, Colorado; 1920. 92 pages.

"Report of a Survey of Public Education in Nassau County, New York." University of the State of New York, Albany, New York; 1918. 283 pages.

SEARS, J. B., et al. "The Boise Survey." World Book Company, Yonkers-on-Hudson, New York; 1920. 290 pages.

SMITH, H. L. "A Survey of a Public School System" (Bloomington, Indiana). Teachers College Contributions to Education, No. 82; 1917. 106 pages.

STRAYER, G. D. "Some Problems in City School Administration" (based on the Butte Survey). World Book Company, Yonkers-on-Hudson, New York; 1916. 225 pages.

TERMAN, L. M. "The Intelligence of School Children." Houghton Mifflin Company, Boston; 1919. 317 pages.

VIRGINIA EDUCATION COMMISSION and VIRGINIA SURVEY STAFF. "Virginia Public Schools." Part One, Reports of Education Commission and Survey Staff (1920), 400 pages; Part Two, Educational Tests (1921), 235 pages. World Book Company, Yonkers-on-Hudson, New York.

WALLIN, J. E. W. "The Mental Health of the School Child." Yale University Press, New Haven, Connecticut; 1914. 450 pages.

WHIPPLE, G. M. "Classes for Gifted Children." Public School Publishing Company, Bloomington, Illinois; 1919. 151 pages.

WOODROW, H. "Brightness and Dullness in Children." J. B. Lippincott Company, Philadelphia; 1919. 322 pages.

II. It is simply impossible to give any comprehensive list of articles from the journals because of the extraordinary amount of material which is being published in this field. But the writers realize that they should not leave the teacher who has become interested in test work without some references which may act as starting points for further reading. So the list below has been included for the benefit of those to whom the back files of educational journals are accessible. The list is barely more than a sampling of articles in the various fields; the effort has been under each topic to choose four or five articles which were either unusually important or typical of this or that special phase of work. If the reader wishes to make an intensive study of any particular type of test he should use these references only as first points of contact, and should then search the current numbers of educational journals — keeping a wary eye for footnotes referring to other articles. He should also feel free to write to some one of the sources of information suggested in Appendix C — or write to the authors of the articles — for further help. Contact with the most recent work can in such ways be readily obtained.

A. Articles Dealing with the Construction and Use of Tests in the School Subjects

Arithmetic

MONROE, W. S. "A Series of Diagnostic Tests in Arithmetic." *Elementary School Journal*, April, 1919.

——"The Derivation of Reasoning Tests in Arithmetic." *School and Society*, September 7 and 14, 1918.

PHILLIPS, F. M. "Value of Daily Drill in Arithmetic." *Journal of Educational Psychology*, Vol. 4, 1913.

SPAULDING, F. T. "An Analysis of the Content of Six Third-Grade Arithmetics." *Journal of Educational Research,* December, 1921.

STONE, C. W. "Standardized Reasoning Tests in Arithmetic." Teachers College Contributions to Education, No. 83, 1916.

WILSON, G. M. "A Survey of the Social and Business Use of Arithmetic." 16th Yearbook of the National Society for the Study of Education, 1917, pages 128–142.

English

AYRES, L. P. "The Measurement of Ability in Spelling." Bulletin of the Division of Education, Russell Sage Foundation, New York, 1915.

COURTIS, S. A. "Uses of the Hillegas Scale." *English Journal,* April, 1919.

DOLCH, E. W. "Measurement of High School English." *Journal of Educational Research,* November, 1921.

FILLERS, H. D. "Oral and Written Errors in Grammar." *Educational Review,* 54:458–470, 1917.

OTIS, A. S. "Reliability of Spelling Scales." *School and Society,* October 18–November 18, 1916.

THEISEN, W. W. "Improving Teachers' Estimates of Composition Specimens with the Aid of the Trabue Nassau County Scale." *School and Society,* February 13, 1918.

TRABUE, M. R. "Supplementing the Hillegas Scale." *Teachers College Record,* January, 1917.

WILLING, M. H. "The Measurement of English Composition in Grades Four to Eight." *English Journal,* March, 1918.

Geography and History

BAGLEY, W. C., and RUGG, H. O. "The Content of American History as Taught in the Seventh and Eighth Grades." University of Illinois School of Education, Bulletin No. 16, 1916.

BELL, J. C., and MCCOLLUM, D. F. "A Study of the Attainments of Pupils in United States History." *Journal of Educational Psychology,* May, 1917.

COURTIS, S. A. "Measuring the Effects of Supervision in Geography." *School and Society,* July 18, 1919.

LACKEY, E. E. "A Scale for Measuring the Ability of Children in Geography." *Journal of Educational Psychology,* May, 1918.

SACKETT, L. W. "A Scale in Ancient History." *Journal of Educational Psychology,* May, 1917.

Reading

BURGESS, M. A. "The Measurement of Silent Reading." Russell Sage Foundation, New York, 1921. 163 pages.

JUDD, C. H. "Reading: Its Nature and Development." Supplementary Educational Monographs, University of Chicago, 1918.

MCCALL, W. A. "A Uniform Method of Scale Construction." *Teachers College Record*, January, 1921.

MONROE, W. S. "Monroe's Standardized Silent Reading Tests." *Journal of Educational Psychology*, January, 1921.

PRESSEY, L. W. "Two Diagnostic Tests in Reading for Use in the Second to Fourth Grades." *Elementary School Journal*, November, 1921.

PRESSEY, L. W., and SKEEL, H. V. "A Group Test for Measuring Reading Vocabulary in the First Grade." *Elementary School Journal*, December, 1920.

PRESSEY, S. L., and L. W. "A Critical Study of the Concept of Silent Reading Ability." *Journal of Educational Psychology*, January, 1921.

THORNDIKE, E. L. "The Measurement of Ability to Read." *Teachers College Record*, September, 1914, and November, 1916.

Writing

AYRES, L. P. "A Scale for Measuring the Handwriting of School Children." Russell Sage Foundation, New York, Bulletin 113, 1915.

BREED, F. S. "The Comparative Accuracy of the Ayres Scale." *Elementary School Journal*, February, 1918.

FREEMAN, F. N. "An Analytical Scale for the Judging of Handwriting." *Elementary School Journal*, April, 1915.

GRAY, C. T. "A Score Card for the Measurement of Handwriting." University of Texas, Bulletin No. 37, 1915.

THORNDIKE, E. L. "Handwriting." *Teachers College Record*, March, 1910.

High School Subjects

ALGEBRA AND GEOMETRY

DOUGLAS, H. R. "A Series of Standardized Diagnostic Tests for the Fundamentals of Elementary Algebra." *Journal of Educational Research*, December, 1921.

MINNICK, J. H. "Minnick's Geometry Scale for Measuring Pupils' Ability to Demonstrate Geometrical Theorems." *School and Society*, February 19, 1919.

MONROE, W. S. "A Test of the Attainment of First-year High School Pupils in Algebra." *School Review*, March, 1915.

RUGG, H. O., and CLARK, J. R. "Standardized Tests and the Improvement of Teaching in First-year Algebra." *School Review*, May, 1917.

LANGUAGES

HANDSCHIN, C. H. "A Test for Discovering Types of Learners in Language Study." *Modern Language Journal*, October, 1918.

HENMON, V. A. C. "The Measurement of Ability in Latin." *Journal of Educational Psychology*, November and December, 1917, and March, 1920.

——"Standardized Vocabulary and Sentence Tests in French." *Journal of Educational Psychology*, February, 1920.

Scales Combining Tests in Several Subjects

BUCKINGHAM, B. R., and MONROE, W. S. "A Testing Program for Elementary Schools." *Journal of Educational Research*, September, 1920.

PINTNER, R. "Results of a Combined Educational and Mental Survey." *Journal of Educational Psychology*, February, 1921.

PRESSEY, L. W. "Scale of Attainment: No. 1." *Journal of Educational Research*, September, 1920.

PRESSEY, S. L. "Scale of Attainment: No. 2." *Journal of Educational Research*, May, 1921.

PRESSEY, L. W. "Scale of Attainment: No. 3." *Journal of Educational Research*, December, 1921.

B. Articles Dealing with the Construction and Use of Tests of Ability

BRACEWELL, R. H. "The Freeman-Rugg Intelligence Tests as an Aid to Economy in School Administration." *School Review*, June, 1921.

BRANSON, N. P. "An Experiment in Arranging High School Sections on the Basis of General Ability." *Journal of Educational Research*, January, 1921.

BRIGHT, I. J. "Intelligence Examinations for High School Freshmen." *Journal of Educational Research*, June, 1921.

CALLIHAN, T. W. "An Experiment in the Use of Intelligence Tests as a Basis for Proper Grouping and Promotions in the Eighth Grade." *Elementary School Journal*, February, 1921.

FORDYCE, C. "Intelligence Tests in Classifying Children in the Elementary School." *Journal of Educational Research*, June, 1921.

HENMON, V. A. C., and STRETZ, R. "A Comparative Study of Four Group Scales for the Primary Grades." *Journal of Educational Research*, March, 1922.

"Intelligence Tests and Their Use." The Twenty-first Yearbook of the National Society for the Study of Education, Parts I and II. Public School Publishing Company, Bloomington, Illinois.

HOLLEY, C. E. "Mental Tests for School Use." Bulletin No. 28, University of Illinois.

LINDSAY, M. D. "Where Test Scores and Teachers' Marks Disagree." *School Review*, November, 1921.

MADSEN, I. N. "Group Intelligence Tests as a Means of Prognosis in High School." *Journal of Educational Research*, January, 1921.

OTIS, A. S. "The Reliability of the Binet Scale and of Pedagogical Scales." *Journal of Educational Research*, September, 1921.

PRESSEY, L. W. "A Scale of Intelligence for Use in the First Three Grades." *Journal of Educational Psychology*, September, 1919.

PRESSEY, S. L. "A Comparison of Two Cities and Their School Systems by Means of a Group Scale of Intelligence." *Educational Administration and Supervision*, February, 1919.

PRESSEY, S. L., and L. W. "Measuring the 'Usefulness' of Tests in Solving School Problems." *School and Society*, November 27, 1920.

C. Articles Dealing with the Organization of Test Work

ALEXANDER, C. "Presenting Educational Measurements so as to Influence the Public Favorably." *Journal of Educational Research*, May, 1921.

BROOKS, F. S. "Standardized Tests in the Rural Schools." *Journal of Educational Research*, May, June, and November, 1920; January, October, November, and December, 1921; February, 1922.

HINES, H. C. "What Los Angeles Is Doing with the Results of Testing." *Journal of Educational Research*, January, 1922.

KEENER, E. E. "Use of Measurements in a Small City School System." *Journal of Educational Research*, March, 1921.

KOOS, F. H. "Educational Measurements in a Small School System." *Journal of Educational Research*, June, 1920.

MADSEN, I. N. "Some Results of a Testing Program in Idaho." *School and Society*, June 1, 1921.

GLOSSARY

ability: from the point of view of education capacity for profiting by instruction, as distinct from skill or knowledge gained from instruction; thus, tests of ability as distinct from tests of achievement.

acceleration: unusually rapid progress through school, usually due to double promotions or extra work in "rapid progress" sections. See *underageness*.

accomplishment quotient: same as achievement quotient.

accuracy score: number of items in an arithmetic test for which the child obtains the correct answer.

achievement: what a child has learned in school to date.

achievement age: the proficiency a pupil has in a particular school subject, expressed in terms of the age of the average child having that degree of proficiency in that subject. So if a child makes a score of 17 on a silent reading test, and 17 is the average score of 11-year-old children, he is said to have an "achievement age" of 11 in silent reading.

achievement quotient: a child's achievement age divided by his mental age; thus, a percentage statement of the extent to which a child has learned his school work in proportion to his ability. If a child's mental age is 10 and his achievement age is 9, his achievement quotient is .90; that is, his work is 90 per cent as good as it should be considering his intelligence.

age-grade table: a tabulation showing the number of children of each age in each grade, as below (the short lines enclose those ages considered "normal" for the grade):

	GRADES						
AGES	1	2	3	4	5	6	TOTAL
5	2						2
6	41	4					45
7	13	22	1				36
8	14	20	11				45
9		7	21	8	1		37
10	2	1	2	22	1		28
11				6	20	7	33
12		1			10	15	26
13						9	9
14					1	2	3
TOTAL PER GRADE	72	55	35	36	33	33	264

In the above table there are 14 eight-year-old children in Grade 1, 20 in Grade 2, 11 in Grade 3. The table may be read either across or up and down. See *mental-age-grade table.*

age norm: the median or average score of a large unselected group of children of a given age.

anatomical age: degree of development of the body structure; usually measured by an X-ray picture of the bones of the wrist showing the degree of ossification.

AQ: achievement quotient or accomplishment quotient.

average score: the sum of the scores divided by the number of scores.

average deviation: the average of all the deviations of a series of measures from the central tendency of that series, regardless of whether the deviations be plus or minus. See *deviation* and *central tendency.*

brightness: the ability of a child as compared with the ability of the average child of that chronological age. So a 4-year-old child with a mental age of 6 is brighter than a 10-year-old child with a mental age of 8, though because of greater maturity the total ability of the second child is greater than the total ability of the first. Brightness is usually expressed in terms of the Intelligence Quotient.

central tendency: either the median, or the average (arithmetical mean).

chronological age: the length of time that has elapsed since a child's birth.

coefficient of correlation: a decimal expressing the degree of relationship between two sets of measures.

comprehension score: number of questions in a reading test that a child answers correctly, or a similar score denoting the child's degree of comprehension of a passage which he has read.

content subjects: such subjects as geography, history, literature, and science, in which knowledge of the content is the chief aim. See *tool subject.*

correct principle score: score given in some tests of problem solving in arithmetic to indicate the number of problems in which the child has used the correct principle, regardless of the correctness of his operations in the fundamentals.

correlation: the relationship between two sets of measures.

correlation table: a "double-entry" table showing the relationship between two sets of measures.

criterion: an independent statement or measure of standing in ability or in achievement in a school subject, by comparison with which a test is judged.

crude score: first results on a test before they have been subjected to any statistical treatment, usually the number of correct answers (also called "raw score").

deviation: the amount by which a score or other measure differs from the average or median of a series of such measures; thus, if a child has a score of 14 and the average is 10, his deviation is plus 4.

diagnosis: a final conclusion regarding the nature and causes of a child's difficulty in school, reached after a consideration of all the known facts.

diagnostic test: a test dealing analytically with specific elements in a school subject, for the purpose of detecting special ability or weakness.

dispersion: the extent to which the scores of a given group differ from one another. See *scatter*.

distribution: a table showing the number of cases at each score, age, grade, or other unit, or interval of units.

educational measurements: tests in the school subjects.

efficiency: the ability of a test to solve, or aid in solving, a practical school problem. Also—a use of the word in an altogether different connection—the ability of a class or pupil in a school subject, as compared with the standard in that subject.

examination: (1) written exercise consisting of questions made out by the teacher; (2) a series of tests combined on a single blank and intended for use together.

fact questions: questions requiring merely knowledge of the facts, as distinct from questions requiring judgment. See *thought questions*.

frequency: the number of cases at each point (as each score) or within each interval on a distribution table.

general test: a test which combines in a single score a statement of total attainment in a school subject or group of subjects, as distinguished from a diagnostic test.

grade norms: the median or average score for children of each grade.

grades: the fundamental divisions of pupils within a school representing different levels of achievement — *not* used to mean the marks given by teachers.

graphical representation: ways of picturing data by means of charts and diagrams, so as to bring out visually the essential facts.

individual differences: differences in ability among children; the demonstration of the extent of such differences is one of the outstanding results of the testing movement.

intelligence: sometimes used to mean general ability, which develops with age; sometimes used to mean brightness, which is presumably constant. The term is therefore ambiguous, unless defined by the user. Used in this book to mean brightness.

intelligence quotient: mental age divided by chronological age, giving a percentage statement of the degree of mental development. In obtaining this statement, both ages should be reduced to months. Thus, if a child has a mental age of 10 yrs., 3 mos., and a chronological age of 12 yrs., 6 mos., his IQ is $\frac{123}{150} = 82$. (Decimal point is disregarded.)

intergrade interval: number of points in score between the median or average score for a grade and the corresponding score for the grade above or below.

interquartile range: the distance from the 25 percentile of a distribution to the 75 percentile. See *percentile.*

IQ: intelligence quotient.

key: a scoring device or list showing the correct answers to a test, and the way in which responses should be scored.

marks: the per cent or other indications ordinarily given by teachers to show the quality of the work of pupils in the school subjects.

median: the middle score of a group of scores when these scores are arranged in order of magnitude. Or, the point on either side of which there are an equal number of scores.

mental ability: innate ability or mental capacity, as distinct from skill or knowledge gained from instruction.

mental age:[1] the mental ability of a child expressed in terms of the

[1] It should be understood by those working with adults that the average mental age of the adult population, though stated by Terman in his account of the Stanford Revision of the Binet Scale to be about 16 years, is now

age of the average child having that ability. Thus, if a child shows a mental ability equal to that of an average 11-year-old child, he is said to have a "mental age" of 11 years.

mental-age-grade table: a table similar to the age-grade table (see *age-grade table*) showing the number of children of each mental age in each grade.

mental maturity: mental adulthood, the maximum mental ability which an individual will attain.

mental measurements: tests of ability, especially of general ability.

minimal essentials: those portions of the field covered by a school subject that are considered absolutely necessary.

motivation: the supplying of a motive or source of effort, which will cause a child to apply himself to his school work.

norm: the median or average performance of children of different ages or grades, as determined by testing large numbers of children.

normal distribution: a symmetrical distribution showing a massing of the cases at the center of the distribution and a gradual decrease in the number of cases toward each extreme — the type of distribution resulting from the operation of chance. When graphed, the curve shows a bell-shaped area. Such a distribution is called normal, because it is the type found most usually in the study of natural and mental phenomena when the study is made from unselected cases.

objective score: an answer which is unmistakably right or wrong, and about which there can be no difference of opinion.

oral reading habits: such habits as whispering the words to oneself, developed in the course of learning to read orally, but detrimental to the development of rapid silent reading.

overageness: (see also *retardation*) a child who is older than the average child in his grade is "overage" for that grade. Usually a

---

believed to be nearer 14, upon evidence gained in the extensive army group testing. The mental ability of the average children of ages above 14 is not known, because so many children drop out of school above 14, and it is impossible to test all the children of any age. So various scales for measuring general intelligence do not give strictly comparable results for ages over 14; and it must not be supposed that a "mental age" of 17, for instance, on a group scale is the same thing as a "mental age" of 17 on the Stanford Binet Scale. The various group scales agree with the Stanford Binet, however, for mental ages of 14 and below.

child of eight or over is considered overage for Grade 1, a child of nine for Grade 2, a child of ten for Grade 3, etc. Retardation is not really synonymous with overageness; overageness is the condition, while retardation is a particular reason for that condition. A child may be overage merely because he entered school late, or for some other reason than because he has been retarded.

overlapping: the extent to which certain children in one group obtain scores the same as those of certain children in another group. So if some sixth-grade children make scores as high as the poorest eighth-grade children, the sixth-grade distribution overlaps the distribution for the eighth grade.

percentile: a division of a distribution according to the per cent of cases falling above or below a given point. So the 10 percentile is the point below which there are 10 per cent of the cases, and above which there are 90 per cent of the cases. See *quartile.*

percentile rank: the percentile at which a given child scores — the percentage of other children in the group scoring below a given child.— As a measure of brightness, the percentage of unselected children of a given age whom a child of that age exceeds in mental ability.

performance tests: tests which measure a child's ability or attainment, not by his ability to give verbal or written answers, but by his ability to do something — solve puzzles, carry out commands, etc.

physiological age: the degree of physiological maturity — maturity of the bodily functions.

practice materials: sets of exercises in a school subject arranged so as to present a consistent series, for the development of ability in that subject.

probable error: (1) that distance above and below the median of a distribution within which 50 per cent of the cases are found, or (2) (as a measure of the reliability of a test) the median error of score, that is, the median number of points by which the actual scores of a test deviate from the corresponding true scores.

P.E.: probable error.

problems: the type of arithmetic question involving reasoning and requiring that the child decide which operations should be used.

prognosis: the forecasting of a child's success or failure in a given line of work. So intelligence tests are frequently used to foretell in grade school a child's probable success in high school, etc.

quality: the degree of general merit in a specimen of handwriting or composition.

quartiles: the 25 and 75 percentiles — that is, the medians of the lower and upper halves of a distribution; the cases halfway between the median and the two extremes.

range: the interval between the highest and lowest scores of a group of children on the same test. See *scatter*.

rate score: the number of problems attempted, the number of letters written, or the number of words read within a specified time.

rating: an estimate made according to a systematized scheme, regarding the ability of an individual in general, or along some one line.

raw score: see *crude score*.

reasoning test: a test in arithmetic involving problem solving.

reliability: the consistency of results yielded by a test; that is, the extent to which the same children tend to get relatively the same score on the same test on successive trials.

reliability coefficient: the coefficient of correlation between successive scores made by pupils if the test is given twice, or a duplicate form is used.

remedial instruction: instruction that seeks to remedy a specific defect; thus, if children cannot "carry," remedial teaching would consist in giving them practice in this very specific detail of arithmetic.

retardation: strictly speaking, slow movement through the grades, as a result of failure and repetition of the work of one or more grades. However, *retardation* is now commonly used to mean the same as *overageness*.

samples: practice exercises to accustom children to the nature of a test.

scale: a test in which the items are arranged in order of increasing difficulty, the increase in difficulty being equal for successive steps.

scatter: the extent to which scores in a group differ from each other. In the first distribution given below, the scatter is wide; in the second distribution there is little scatter. Practically synonymous with *dispersion*.

Scores	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Dis. 1		1	4	5	2	6	7	4	3	2	1		2	1	1	No. cases — 39
Dis. 2							10	11	14	9	11	10				" " — 65

scatter diagram: same as correlation table, showing the scatter for two distributions at the same time and the relation between the two measures.

school census: a census of all the children of school age within a district or community.

score: the final statement of a child's standing on a test. See *crude score.*

silent reading: reading without vocalization — assimilative reading.

specimen: a sample, as of handwriting or composition, obtained for rating by comparison with the type specimens of a scale.

spelling demons: common words that are particularly difficult to spell, as "necessary" or "receive."

standard: the proficiency in a given subject which should be reached; the level of achievement in a given subject to which a school should aim to bring its children. Thus children should learn to write a quality of 60 in the Ayres Scale, since this quality is needed in the business world; quality 60 then becomes a "standard." Again, if a goal for fourth-grade pupils is set at 20 examples in addition of a certain type to be done in 5 minutes, this score is called a standard score.

standardized tests: tests which are carefully worked out as to procedure in giving and scoring, and which have been given to enough children so that reliable norms have been obtained.

survey: an extensive investigation of an entire school or system in which tests are usually an important means of investigation.

tabulation: a table upon which the score of each child is represented by a mark; the table thus gathers together the results and presents them as a unified whole.

teachers' estimates: opinions of teachers concerning their pupils based upon their acquaintance with the pupils and *not* on test results.

thought questions: questions requiring judgment on the part of the child, as distinct from those requiring only knowledge of facts.

timed sentence spelling tests: spelling tests in which the words to be spelled are arranged in sentences which are dictated to the children at a given rate, thus requiring them to spell the words as they would probably spell them when their minds were upon the context of what they were writing rather than upon the spelling of an isolated word.

tool subjects: subjects such as reading, writing, and the fundamental processes of arithmetic which are used as tools in the mastery

of such further subjects as history, science, etc., and in the ordinary demands of later life.

underageness (commonly used to mean the same as *acceleration*): a child who is underage is one who is younger than the average child in his grade. Underageness is a condition of which acceleration is only a cause (see *overageness*). A child 5 years old or younger is considered underage for Grade 1, 6 years old or younger for Grade 2, 7 years old or younger for Grade 3, etc.

validity: the extent to which a test really measures the ability it is supposed to measure.

variability: the extent to which the same child tends to vary in his performance on the same test at different times. Also, the extent to which the scores of a group of children differ from each other. See *dispersion* and *scatter*.

weighting: the assigning of special values to a test or item because of its special significance.

INDEX

(For definition and explanation of terms see Glossary.)